Journal of Biomedical Informatics xxx (2014) xxx–xxx

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Application of time series discretization using evolutionary programming for classification of precancerous cervical lesions

Héctor-Gabriel Acosta-Mesa [a,*], Fernando Rechy-Ramírez [a], Efrén Mezura-Montes [a], Nicandro Cruz-Ramírez [a], Rodolfo Hernández Jiménez [b]

[a] School of Physics and Artificial Intelligence, Department of Artificial Intelligence, Universidad Veracruzana, Sebastián Camacho # 5, 91000 Xalapa, Veracruz, Mexico
[b] Obstetrician and Gynaecologist, Diego Leño # 22, C.P. 91000 Xalapa, Veracruz, Mexico

ARTICLE INFO

Article history:
Received 15 June 2013
Accepted 3 March 2014
Available online xxxx

Keywords:
Times series discretization
Evolutionary algorithms
Classification
Cervical cancer detection

ABSTRACT

In this work, we present a novel application of time series discretization using evolutionary programming for the classification of precancerous cervical lesions. The approach optimizes the number of intervals in which the length and amplitude of the time series should be compressed, preserving the important information for classification purposes. Using evolutionary programming, the search for a good discretization scheme is guided by a cost function which considers three criteria: the entropy regarding the classification, the complexity measured as the number of different strings needed to represent the complete data set, and the compression rate assessed as the length of the discrete representation. This discretization approach is evaluated using a time series data based on temporal patterns observed during a classical test used in cervical cancer detection; the classification accuracy reached by our method is compared with the well-known times series discretization algorithm SAX and the dimensionality reduction method PCA. Statistical analysis of the classification accuracy shows that the discrete representation is as efficient as the complete raw representation for the present application, reducing the dimensionality of the time series length by 97%. This representation is also very competitive in terms of classification accuracy when compared with similar approaches.

1. Introduction

Many real-world applications related with information processing generate temporal data [14]. The temporal databases generated require enormous data storage. It is therefore desirable to compress this information while maintaining the most informative features. Previous work on this topic has been mainly focused on data compression. However, they do not rely on significant information measured with entropy [13,15]. In those approaches, the dimensionality reduction is given by the transformation of time series of length N into a dataset of n coefficients, where $n < N$ [10]. The two main characteristics of a time series discretization scheme are as follows: the number of segments in which the time series length has to be partitioned (word size) and the number of intervals required to represent its amplitude expressed by continuous values (alphabet). Fig. 1 shows a time series with a grid that represents the cut points for word size = 9 and alphabet = 7. Using this transformation each time series is discretized and represented as a string.

Among the approaches proposed to deal with time series data discretization we find those which work with one time series at a time, such as the one proposed by Mörchen and Ultsch [17]. His algorithm is centered on the search of persistent states (the most frequent values) in time series. However, such states are not common in many real-world applications for time series. Another representative approach was proposed by Dimitrova et al. [7], where a multi-connected graph representation for time series was employed. The links between nodes have Euclidean distance values which are used under this representation to eliminate links in order to obtain a path that defines the discretization scheme. Nonetheless, this way to define the discretization process could be disadvantageous because not all the time series in a data set will necessarily have the same discretization scheme.

Keogh et al. [15] proposed the Symbolic Aggregate Approximation (SAX) approach. This algorithm is based on the Piecewise Aggregate Approximation (PAA), a dimensionality reduction algorithm [11]. After PAA is applied, the values are then trans-

* Corresponding author. Address: School of Physics and Artificial Intelligence, Department of Artificial Intelligence, Universidad Veracruzana, Sebastian Camacho 5, Col. Centro, C.P. 91000 Xalapa, Veracruz, Mexico. Fax: +52 552288172957.
E-mail addresses: heacosta@uv.mx (H.-G. Acosta-Mesa), frechyr@hotmail.com (F. Rechy-Ramírez), emezura@uv.mx (E. Mezura-Montes), ncruz@uv.mx (N. Cruz-Ramírez), roheji@msn.com (R. Hernández Jiménez).

http://dx.doi.org/10.1016/j.jbi.2014.03.004
1532-0464/

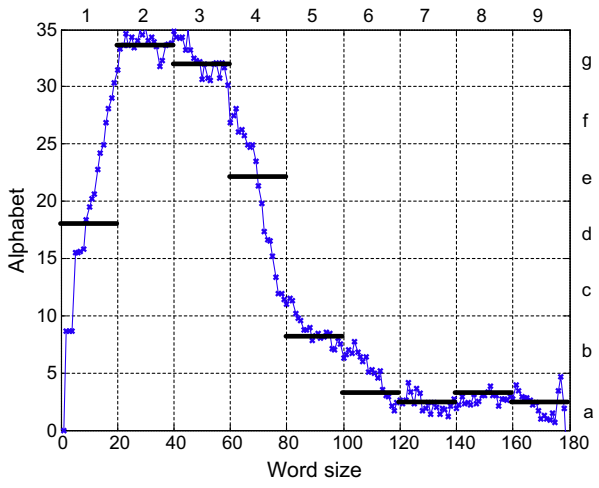

Fig. 1. Time series discrete representation. The time series length has been arbitrarily divided in nine segments of equal length. The continuous values of the time series amplitude are discretized in seven intervals of equal length as well. In this case the time series is represented as the string: dggebaaaa.

formed into categorical values through a probability distribution function. Although SAX is an improvement of PAA, both algorithms require the alphabet and the word size as inputs, which is their main disadvantage because it is not clear how to define them for a given time series dataset.

There are other approaches based on search algorithms, for example García-Lopez et al. [6] proposed EBLA2, which in order to automatically find the word size and alphabet performed a greedy search looking for entropy minimization. The main disadvantage of this approach is the sensitivity of the greedy search leads it to get trapped in local optima. Therefore, they used simulated annealing as a search algorithm and the results were improved. Finally, in [1], a genetic algorithm was used to guide the search; however the solution was incomplete in the sense that the algorithm considered the minimization of the alphabet and the word size as two sequential and independent processes. In this way some solutions could not be generated and the obtained solution is not global. In order to avoid it, we developed an algorithm that automatically finds both parameters at the same time [5].

This is the most important contribution of our approach, since most of the discretization algorithms require, as an input, the parameters of word size and alphabet [12,15]. However, in real-world applications it might be very difficult to know in advance their best values. Hence, their definitions require a careful analysis of the time series data set. In this work, we introduce the main ideas behind our discretization approach in which both the word size and the alphabet are calculated automatically, and how this approach can be applied to the medical field in the classification of precancerous cervical lesions.

Cervical cancer is the second leading cause of death for women worldwide. If it is detected early, the probability of cure is very high. After Pap smear test, colposcopy is the most used technique to diagnose this disease due to its higher sensitivity and specificity. Colposcopy allows us to visualize the uterine cervix using a microscope fitted with a light source. During the colposcopic test, the appearance of the cervix is observed while a solution of acetic acid is spread on the epithelium, which produces a change from the usual pink tissue to a whitish color due to the coagulation of proteins in the cellular nucleus. This phenomenon is called acetowhitening, and its effect is more evident at the wavelet of 525 ± 15 (green) due to the hemoglobin absorption made by the stroma. Acetowhitening disappears in less than 10 min and is more evident in precancerous lesions due to an altered nucleus to cytoplasm ratio (Fig. 2).

The primary problem with this technique is the intrinsic subjectivity of the test, i.e., the amount and speed of color change perceived could be different for various observers; this fact may produce high variability on the diagnoses made by experts. Therefore changes in the mechanisms to quantify the amount of acetowhite change and the speed changes are needed to improve the test. Some researchers have suggested using the temporal patterns intrinsic to the color changes, which we called Aceto-White Response Functions (AWRF).

Costas J. Balas et al. proposed the use of spectroscopy to study the correlation between aceto-white patterns and precancerous cervical lesions [2]. Although some efforts have been made to characterize precancerous cervical lesions using aceto-white temporal patterns, to the best of our knowledge there is not a complete understanding of how to automatically analyze colposcopic images using aceto-white temporal patterns for classification of cervical tissue. In our previous work [3,4], we compared the shape of the temporal patterns to establish relationships among similar shapes, and the correlation of those patterns with certain types of tissue.

In the present work we apply to these temporal patterns our discretization approach in order to compress and analyze the structural properties of the intrinsic dynamics involved in the aceto-white phenomenon. In order to find a competitive discretization scheme that provides a suitable word size and alphabet, and considering its simplicity with respect to other evolutionary algorithms, evolutionary programming (EP) is adopted as a search algorithm: no recombination and parent selection mechanisms are performed and just mutation and replacement need to be designed. In a previous work the efficiency of this approach was assessed using 20 time series databases of the UCR Time Series Classification/Clustering repository [5].

The main contribution of this work is the application of our discretization approach to a medical domain in order to contribute to the solution of one of the most important health problems. The contents of this paper are organized as follows: Section 2 introduces the context in which the application is motivated. After that, Section 3 exposes the main ideas behind the proposed discretization approach. Section 4 presents results. Finally, Section 5 draws some conclusions and prospects for future work.

2. Materials

2.1. Data acquisition

Two hundred women were included in this study, within the total number of patients, in 100 cases a tissue sample or biopsy was obtained because some changes in the appearance of the cervical epithelium were observed by the colposcopist and these alterations led to suspicion of a lesion. In the other 100 patients, the specialist did not find changes that suggested the presence of a lesion and due to clinical protocols a biopsy was not taken. Of the total quota, 93 cases were positive for precursor lesions of cervical cancer and 107 negative. Before the test, the patients signed an informed consent.

Subsequently, during the colposcopic test a set of digital images were obtained. The acquisition was performed using a colposcope Vasconcellos CP-M1225 with an STC-N63BJ camera. Because in previous research [2] it has been reported that the acetowhitening effect is higher at the wavelet of 525 ± 15 (green), and the image acquisition was made using a green optical filter. The dimension of the images was 352 × 240 pixels with a sampling frequency of 1 frame/second. The images were stored as separated files in the BMP format. A tool was developed for the acquisition and it was implemented in MATLAB 7.0. Before the application of three

Please cite this article in press as: Acosta-Mesa H-G et al. Application of time series discretization using evolutionary programming for classification of precancerous cervical lesions. J Biomed Inform (2014), http://dx.doi.org/10.1016/j.jbi.2014.03.004