clustering, fuzzy number,
information granulation, interpolation,
reconstruction criterion

Michał MOMOT [1], Alina MOMOT [2], Adam GACEK[3]

GRANULAR REPRESENTATION OF BIOMEDICAL SIGNALS USING NUMERICAL DIFFERENTIATION METHODS

This work presents the general idea of granular description of temporal signal, particularly biomedical signal sampled with constant frequency. The main idea of presented method is based on using triangular fuzzy numbers as information granules in temporal and amplitude spaces. The amplitude space contains values of first few derivatives of underlying signal. The construction of data granules is performed using the optimization method according to some objective function, which balances the high coverage ability and the low support of fuzzy numbers. The granules (descriptors) undergo the clustering process, namely fuzzy c-means. The centroids of created clusters form a granular vocabulary and the quality of description is quantitatively assessed by reconstruction criterion.

There are presented results of experiments with the electrocardiographic signal, digitally sampled and stored in MIT-BIH database. The method of numerical differentiation of function based on finite set of its values is employed, which incorporates polynomial interpolation. The paper presents results of numerical experiments which show the impact of method parameters, such as temporal window length, degree of polynomial, fuzzification parameter, on the reconstruction ability of presented method.

1. INTRODUCTION

The general concept of information granulation and granular computing in particular plays important role in many human-centered approaches to data processing [1]. There is a vast set of scientific works in this area. Among many applications of data granulation there exists an idea of using fuzzy sets, especially fuzzy numbers [11], to construct the granules (descriptors) for the purpose of temporal signal processing. In [5] there is presented the application of fuzzy clustering algorithm to build a certain semantics of descriptors in form of triangular or trapezoidal fuzzy numbers. The parameters of these numbers, being data granules, were set according to some optimization criterion. As a result there was presented the ability to describe some class of temporal signals, namely electrocardiographic signal, in a digital form together with quantitative assessment of description quality. The granular reconstruction of examined signal was partly based on its first derivative, which reflected the changes of signal level. However, the most popular simple method of numerical computing estimate of signal derivative, sampled with finite frequency (i.e. one-step difference), is usually highly vulnerable to distort, because of the presence of noise. There exist various differentiation schemes, which allow to estimate the function derivatives based on partial knowledge i.e. finite set of function values [3]. One of the primary objectives in constructing such schemes is their robustness to input errors (noises), which inherently accompany useful part of signal. The other objective is low computational complexity, which enables the method to operate in a real-time regime.

The aim of this paper is to describe the method of transforming temporal data, particularly biomedical signal, into the form of data granules set. The presented approach is similar to one previously presented in [5]. However, some significant modifications have been made: using only relative values of signal amplitude (which was achieved by using first few derivatives of signal), using numerical differentiation scheme based on interpolation polynomial (for the purpose of higher robustness to noise). Another significant assumption was made, as to avoidance of dependence on any prior information about

[1] Institute of Medical Technology and Equipment ITAM, Roosevelt'a 118 Street, 41-800 Zabrze, michal.momot@itam.zabrze.pl
[2] Silesian University of Technology, Faculty Of Automatic Control, Electronics And Computer Science, Akademicka 16 Street, 44-100 Gliwice, alina.momot.@polsl.pl
[3] Institute of Medical Technology and Equipment ITAM, Roosevelt'a 118 Street, 41-800 Zabrze, adam.gacek@itam.zabrze.pl

underlying signal. For instance, the presented method does not require knowledge about temporal or amplitude markers (characteristic points) to operate.

The latter part of article is organized according to the following plan. Section 2 contains description of presented method, including the main idea of constructing data granules, algorithm of numerical differentiation and also the creating granular vocabulary by using clustering algorithm together with definition of reconstruction quality measure. Section 3 contains description of numerical experiments, discussion of results obtained and some implementation remarks. This section also contains a discussion of possible future plans for the development of the presented method.

2. METHOD DESCRIPTION

2.1. CONSTRUCTING OF DATA GRANULES

The temporal character of proposed method consists in splitting underlying signal into non-overlapping segments (time windows) with fixed length L samples, i.e. $[s_{iL}, s_{iL+1}, \ldots, s_{i(L+1)-1}]$. For ith segment the data granule is constructed as a triangular fuzzy number based on the values of signal amplitude (or rather, its derivatives), therefore this part of granulation process has spatial nature. The symmetric triangular fuzzy numbers were chosen because of their unimodal membership function:

$$\mu_{c,r}(x) = \max\left\{1 - \frac{|x-c|}{r}, 0\right\}, \tag{1}$$

where c is center of the number and r is its radius. Another justification for using this type of fuzzy numbers is the concise description of generated data granules, because every granule requires $2K$ real numbers, when the derivatives up to order K are under consideration. Figure 1 illustrates construction of data granule for first order derivative of exemplary signal.

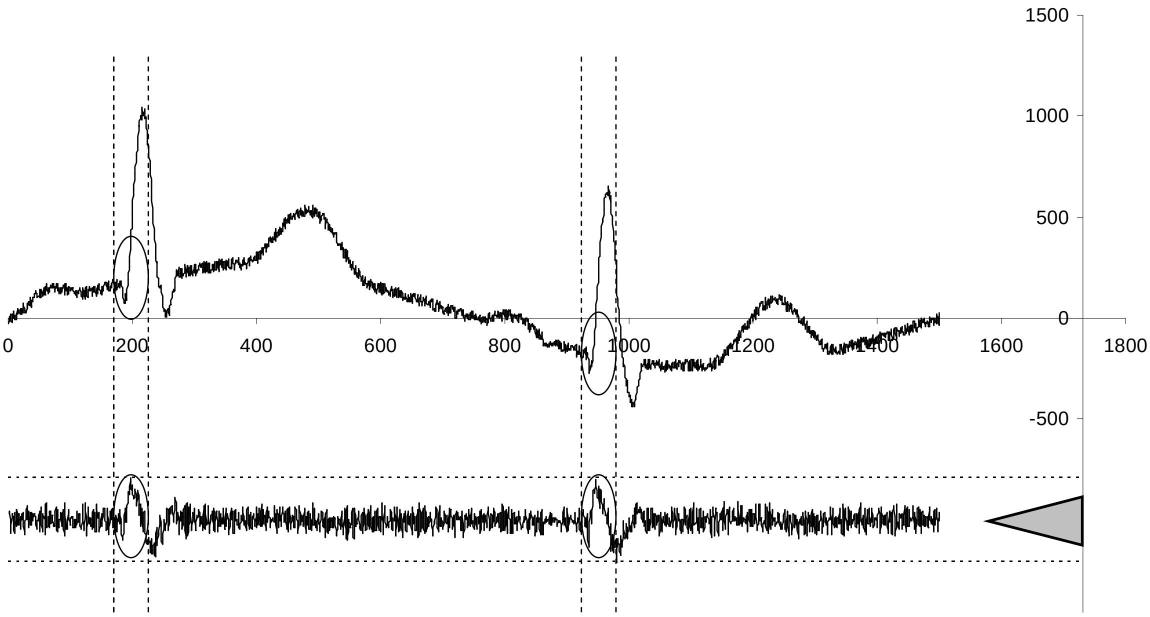

Fig. 1. Illustration of construction of single data granule for first order derivative of exemplary signal.

The center of each fuzzy number is equal to median of values in corresponding segment. The choice of median was justified by its high tolerance to outliers in the data. Assumption has been made that the length of time window is an odd number, therefore

$$c_i^{(k)} = \tilde{s}^{(k)}_{iL+\frac{L-1}{2}}, \tag{2}$$

where $c_i^{(k)}$ is the center of fuzzy number, $[\tilde{s}^{(k)}_{iL}, \tilde{s}^{(k)}_{iL+1}, \ldots, \tilde{s}^{(k)}_{i(L+1)-1}]$ is the vector of signal derivatives sorted in ascending order and *k* is the order of derivative.

The choice of fuzzy number radius is realized by accommodating two requirements. First, the empirical evidence of granule should be as high as possible, which implies maximizing the overall membership for data segment:

$$\sum_{j=0}^{L-1} \mu_{c_i^{(k)},r}\left(\tilde{s}^{(k)}_{iL+j}\right). \tag{3}$$

The second objective states that the granule should be maximum specific, which is realized by minimizing the radius of corresponding triangular fuzzy number. These two conflicting requirements lead to following optimization rule:

$$r_i^{(k)} = \arg\max_{r\in(0,+\infty)} \frac{\sum_{j=0}^{L-1} \mu_{c_i^{(k)},r}\left(\tilde{s}^{(k)}_{iL+j}\right)}{r}. \tag{4}$$

Applying above criterions results in following granular description:

$$[c_i^{(1)}, r_i^{(1)}, \ldots, c_i^{(K)}, r_i^{(K)}] \tag{5}$$

for the *i*th segment and *K* first derivatives of the underlying signal.

2.2. NUMERICAL DIFFERENTIATION SCHEME

Using signal derivatives instead of actual samples is motivated by the need of making the method independent on constant or slowly varying components of signal, which enables the granules to capture the actual nature of analyzed data segments. For example, in electrocardiographic signal there is well known phenomenon of base-line wander (slow changes of isoelectric line), see Figure 1. There are several approaches to overcome this problem, including base-line shifting, normalization (with constant or adaptively varying scaling factors) [10] or using derivatives of observed signal. In case of derivatives, the simplest idea is to use one-step difference:

$$s_n^{(1)} = s_{n+1} - s_n \tag{6}$$

or its symmetric version:

$$s_n^{(1)} = \frac{s_{n+1} - s_{n-1}}{2}. \tag{7}$$

However, these methods are very sensitive to the presence of signal disturbances, especially high-frequency noise. To overcome this problem more sophisticated algorithms can be applied [3][4], including differential quadratures based on polynomial interpolation [12]. The proposed signal granulation method exploits Lagrange polynomial interpolation with equidistant nodes within symmetric fixed-width time window. The derivative of polynomial is taken as an estimate of signal derivative. The algorithm of computing coefficients of this polynomial is briefly described below.

Let R denote the radius of time windows around the Nth sample of signal. The values of interpolated function form a vector $\mathbf{s} = [s_{N-R}, \ldots, s_{N-1}, s_N, s_{N+1}, \ldots, s_{N+R}]^T$, the degree of polynomial g is determined by number of samples and is equal to $2R$, therefore

$$g(x) = \sum_{j=0}^{2R} a_j x^j \tag{8}$$

and it is equivalent to vector $\mathbf{a} = [a_0, a_1, \ldots, a_{2R-1}, a_{2R}]^T$. Since the nodes of interpolation are equally distributed in time window (in fact each node corresponds to one sample index), it can be assumed, without any loss of generality, that the nodes are of the form:

$$x_{N-R} = -R, \ldots, x_{N-1} = -1, x_N = 0, x_{N+1} = 1, \ldots, x_{N+R} = R, \tag{9}$$

the interpolation procedure is time-invariant and depends only on values of interpolated function. Vector of polynomial coefficients may be obtained by solving the system of linear equations:

$$\mathbf{Va} = \mathbf{s}, \tag{10}$$

where $(\mathbf{V})_{i,j} = i^j$ for $i = -R, \ldots, R$, $j = 0, \ldots, 2R$, i.e.

$$\mathbf{V} = \begin{bmatrix} 1 & -R & \cdots & (-R)^{2R-1} & (-R)^{2R} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ -1 & 1 & \cdots & -1 & 1 \\ 1 & 0 & \cdots & 0 & 0 \\ 1 & 1 & \cdots & 1 & 1 \\ \vdots & \vdots & \cdot^{\cdot^{\cdot}} & \vdots & \vdots \\ 1 & R & \cdots & R^{2R-1} & R^{2R} \end{bmatrix}. \tag{10}$$

The matrix $\mathbf{V}$ is non-singular, hence $\mathbf{a} = \mathbf{V}^{-1}\mathbf{s}$. Moreover, there is no need to explicitly determine all polynomial coefficients, since $s_N^{(1)} = g'(0) = a_1 = \mathbf{d}_1^T \mathbf{V}^{-1}\mathbf{s}$ where $\mathbf{d}_1 = [0,1,0,\ldots,0]^T$. Thus the differentiation procedure may be seen as high-pass filtering using FIR filter with constant coefficients given by the formula $\mathbf{d}_1^T \mathbf{V}^{-1}$. The higher order derivatives are calculated in a similar manner.

2.3. CREATING GRANULAR VOCABULARY USING FUZZY CLUSTERING

The granular representation of data segments, being collection of $2K$-dimensional vectors is subject to clustering process, which leads to determining descriptors that reflect the structure of whole set of granules. In proposed approach the well known fuzzy c-means (FCM) method is used [2][7]. The main idea of FCM is to determine, given the collection of N input data in form of D-dimensional vectors, the membership values $u_{i,k}$ that minimize the following objective function

$$Q = \sum_{i=1}^{c} \sum_{k=1}^{N} u_{i,k}^m \left\| \mathbf{z}_k - \mathbf{v}_i \right\|^2, \tag{11}$$

where $\mathbf{v}_i$ are cluster prototypes (centroids), $\|\cdot\|$ denotes some norm in D-dimensional space (usually Euclidean norm) and $m > 1$ is fuzziness parameter controlling impact of the membership grades on individual clusters. The clustering procedure is performed in an iterative manner, alternating between computing the membership values and updating cluster prototypes, until some termination condition is fulfilled. After determining cluster prototypes coordinates, the following formula enables to reconstruct input data vector based on obtained granular vocabulary:

$$h(\mathbf{z}) = \frac{\sum_{i=1}^{c} \left(u_i(\mathbf{z})\right)^m \mathbf{v}_i}{\sum_{i=1}^{c} \left(u_i(\mathbf{z})\right)^m}, \tag{12}$$

where $u_i(\mathbf{z})$ is the degree of membership in ith cluster for the given input vector $\mathbf{z}$. This leads to the reconstruction criterion

$$V = \frac{1}{N} \sum_{k=1}^{N} \left\| \mathbf{z}_k - h(\mathbf{z}_k) \right\|^2, \tag{13}$$

which expresses average error of granular vocabulary explanation for the set of data. It might serve as a quantitative measure of method performance. It should be stressed that the value of V depends on method parameters, namely granulation window length L, maximum order of derivative K, interpolation window radius R, as well as parameters controlling clustering process: clusters number c and fuzziness parameter m. Therefore the question of interest is how to choose suitable values of these parameters in order to obtain best performance of the proposed method.

3. RESULTS AND DISCUSSION

The numerical experiment was performed in order to experimentally evaluate performance of proposed method. The input data were a part of MIT-BIH database [8, [9]. This set of data contains excerpts from two-channel ambulatory ECG recordings. For the purpose of this study over 10 minutes of signal have been considered and the recordings came from different patients. Only single channel of signal was taken into account and the tests did not incorporate any prior information, like human- or computer-readable annotations. Derivatives of first and second order were calculated (K=2) with the interpolation window radius R=2 (on the basis of visual assessment of the quality of the signal derivative). The granulation window length L varied from 5 to 11 (only odd numbers) and the maximum number of clusters was set to 6. The input values of signal, as well as results of computation were expressed in μV.

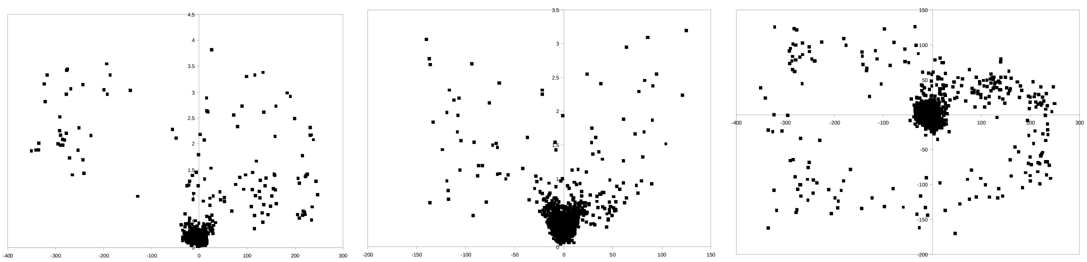

Fig. 2. Illustration of centers and radii of fuzzy numbers for 1st (left), 2nd (middle) derivative and centers only for both derivatives (right) of exemplary signal scattered over the plane.

Figure 2 graphically presents the exemplary structure of obtained data granules by means of centers and radii of determined triangular fuzzy numbers for first two derivatives (for L=5). The figure reveals visible regularities, however the structure is fully described by 4-dimensional vectors and it can not be fully assessed by optical examination of scatter plots. The clustering procedure was performed and the reconstruction ability was empirically evaluated. For every pair of L and c the empirical reconstruction error was presented as a function of fuzziness parameter m (in range 1-3). The optimal values m_{opt} of parameter m was determined, which resulted in minimal reconstruction error V_{opt}. Figure 3 presents reconstruction error as a function of parameter m for c=2,…,6 and L=5 (on the left), L=11 (on the right). In general, increasing c and L results in better reconstruction ability. The charts reveal also the interesting phenomenon: in most cases the reconstruction error increases rapidly for m greater than some threshold value m_{thresh}. However, for L>7 and c>3 the functions become more smooth and the point of rapid change moves to the right (m_{thresh}>2). It suggests that the method will work stable for selected values of parameter m, namely in a range 1.0<m<2.0. Table 1 presents exact empirical values of m_{opt}, V_{opt} and m_{thresh}.

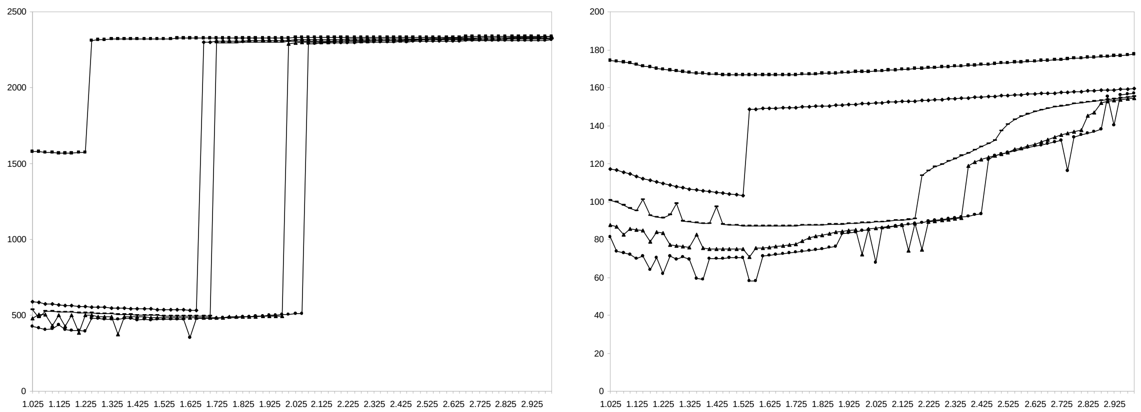

Fig. 3. Reconstruction criterion as a function of parameter m for c=2,…,6 and L=5 (left), L=11 (right).

Table 1. Optimal values of reconstruction criterion and corresponding parameters.

L	c	V_{opt}	m_{opt}	m_{thresh}
5	2	1569.15	1.150	1.225
	3	533.28	1.650	1.650
	4	484.26	1.050	1.725
	5	374.97	1.350	1.975
	6	350.20	1.625	2.050
7	2	944.37	1.175	1.325
	3	311.03	1.775	1.775
	4	285.93	1.725	2.000
	5	235.96	1.500	2.100
	6	209.52	1.325	2.200
9	2	432.81	1.300	1.900
	3	201.48	1.900	1.925
	4	167.19	1.225	2.575
	5	148.65	1.150	2.750
	6	116.89	1.450	2.875
11	2	166.63	1.575	-
	3	102.27	1.700	1.525
	4	87.16	1.625	2.175
	5	70.89	1.550	2.350
	6	58.13	1.550	2.425

It is worth making a mention about how the described method has been implemented. Since the presented algorithms have high computational complexity, there is a need for techniques to reduce computation time. It can be easily observed that processes of constructing data granules in separate granulation windows are performed independently from each other, therefore they can be performed concurrently. The use of technology CUDA (Compute Unified Device Architecture, [6]) at this stage made it possible to significantly reduce computation time.

The issues described in this article will be the subject of further research. It is planned to develop a modified and extended version of the presented method, especially considering other types of fuzzy numbers, the modified optimality criteria for the creation of data granules, as well as other clustering algorithms. Some other plans relate to applications of described method. Since, as has been mentioned previously, the granulation method itself does not require any prior information about the characteristic points of analyzed signal, it is possible to apply it to the calculation of linguistic description function of biomedical signals.

ACKNOWLEDGEMENTS

This research was partially supported by Polish Ministry of Science and Higher Education as Research Project N N518 406438.

BIBLIOGRAPHY

[1] BARGIELA A., PEDRYCZ W., Granular Computing: An Introduction, Kluwer Academic Publishers, Dordercht, 2003.
[2] BEZDEK J., Pattern Recognition with Fuzzy Objective Function Algorithms, Plenum Press, New York, 1981.
[3] BRACI M., DIOP S., On numerical differentiation algorithms for nonlinear estimation, Proc. 42nd IEEE Conf. on Decision and Control, Maui, Hawaii USA, 2003, pp. 2896-2901.
[4] BURDEN R.L., FAIRES J.D., Numerical Analysis, Brooks/Cole, 2000.
[5] GACEK A., PEDRYCZ W., A Granular Description of ECG Signals, IEEE Trans. Biomed. Eng., Vol. 53, No. 10, 2006, pp. 1972-1982.
[6] GARLAND M. et al., Parallel Computing Experiences with CUDA, IEEE Micro, Vol. 28, No. 4, 2008, pp. 13–27.
[7] HATHAWAY R., BEZDEK J., HU Y., Generalized fuzzy c-means clustering strategies using L_p norm distances, IEEE Trans. Fuzzy Syst., Vol. 8, No. 5, 2000, pp. 576-582.
[8] MARK R., MOODY G., MIT-BIH Arrhythmia Database Directory, MIT, Cambridge, MA, 1988.
[9] MOODY G. B., MARK R. G., The MIT-BIH arrhythmia database on CD-ROM and software for use with it, Proc. Conf. Computers in Cardiology, San Diego, CA, 1990, pp. 185-188.
[10] ORTOLANI M., HOFER H., et al., Fuzzy Information Granules in Time Series Data, Proc. IEEE Int. Conf. on Fuzzy Systems, 2002, pp. 695-699.
[11] PEDRYCZ W., Fuzzy Sets as a User-Centric Processing Framework of Granular Computing. In: Pedrycz W., Skowron A. and Kreinovich V (eds), Handbook of Granular Computing, John Wiley & Sons, Chichester, 2008, pp. 97-139.
[12] ZHANG Y., Advanced Differential Quadrature Methods, CRC Press, Boca Raton, 2009.