CMST 19(2) 99-105 (2013) DOI:10.12921/cmst.2013.19.02.99-105

A Representative Set Method for Symbolic Sequence Clustering

B. Kozarzewski

University of Information Technology and Management
ul. H. Sucharskiego 2, 35-225 Rzeszów, Poland
E-mail: bkozarzewski@wsiz.rzeszow.pl

Received: 25 October 2012; revised: 24 February 2013; accepted: 1 March 2013; published online: 22 May 2013

Abstract: Sequence decomposition into a set of consecutive, distinct subsequences is crucial for symbolic sequence analysis. It reduces significantly the reference base of the recorded sequence for further retrieval and allows for original similarity and membership measures of the sequences. The introduced measures are a start point to a new algorithm for clustering sequences into groups of similar individuals. Algorithms that use the concept of a representative set achieved relatively good clustering results. The representative set that we have introduced is precisely and uniquely defined in contrast to that used in other applications.

Key words: similarity and membership measures; representative set; clustering

I. INTRODUCTION

Object clustering is a method of grouping individuals that are somehow similar. The widely used approach consists in computing pair wise similarities and next clustering the sequences by using them. A certain of measure that can determine whether and how similar two individuals are is required. Most clustering methods use various distance measures. An alternative to the distance measure is the similarity function. An example is the cosine of the angle between two vectors representing objects. The distance measure requires a geometrical representation of an object. This, however, is far from being unique in the case of a symbolic sequence [1, 2]. There are several definitions of similarity as well [3-5]. The similarity measure we will use does not require geometrical representation of sequences. It is based on the common number of distinct subsequences, (in general, an ordered list of elements: tuples, mers or words) which are members of both sequences of interest. So is a membership of the sequence in a set of sequences, its measure is a ratio of subsequences which are in common with the set.

There are numerous clustering methods that are in use [6-8]. They can be roughly divided into hierarchical clustering and partition clustering methods. In the hierarchical clustering, the number of clusters need not be known in the beginning. The hierarchy is built up in a series of steps starting with each object as an individual cluster. The partitioning method arranges all objects into predefined groups, each object belonging to one group (cluster). The expected result of such a process is typology, with each cluster grouping individuals similar to each other. Once a set of clusters has been identified, a new object has to be compared against a collection of sequences of all clusters before it is assigned to a particular cluster. The corresponding algorithm has an $\mathcal{O}(n^2)$ computational complexity in the size n of the collection. In this work we present a method for sequence clustering that is based on a representative set of mers. The method follows the idea of extracting the key features of each cluster to define a kind of fingerprint identifying a collection of categorical sequences – candidates for a certain cluster [9-10]. However, instead choosing a part of each sequence as the member of representative, we propose a set of mers that share

large enough portions of mers of the sequence collection (union set of mers representing the selected collection) as the fingerprint mentioned. We call it the representative set of the cluster. The computational complexity of the sequences assigning algorithm falls then to $\mathcal{O}(n)$.

The paper is organized as follows: in the second section the representative set (RS in short), similarity and membership measures of symbolic sequences are defined and discussed. As an illustration, sets of mitochondrial genomes of birds, fishes, mammals and turtles are analyzed in section 3. Size (number of mers) of RS of mammals versus RS coverage (number of sequences whose spectra enter the RS) is discussed. Representativeness (average membership of other sequences of the group in RS) versus RS coverage is derived. Then the following three methods of sequence clustering are presented:

a) Typical agglomerative similarity-based clustering with a discussion of an unavoidable problem when to stop the clustering process and determine the correct number of clusters [11]
b) Agglomerative clustering based on representative sets selected from candidates of the lowest variance
c) Partitioning method based on RS sets with a number of cluster set in advance.

The last section contains the final summary comments.

II. SIMILARITY, REPRESENTATIVE SET AND MEMBERSHIP MEASURE

Very long sequences, which occur in many fields, when directly represented by vectors are usually difficult to analyse or compare. Examples are genomes or share prices. Decreasing the size of a vector before the actual object grouping has numerous advantages. Mechanical decomposition of a sequence into a set of short subsequences (called k-mers) makes sense for statistical analysis only. Subsequences of the same physical meaning may differ in length and can exist in different places of the sequence. An answer to the question how to perform meaningful sequence decomposition is to use a modified Ke and Tong algorithm [12]. The result of the decomposition is a set of ordered, distinct and non-overlapping subsequences, representing a sequence of symbols. The collection of all mers will be called in the following as the spectrum of the primary symbolic sequence. In [12] the total number of mers was considered as the main result of the decomposition procedure and considered as a measure of complexity of the sequence. In [13] it was shown that the spectrum appears to be a very rich resource of information on the symbolic sequence.

Similarity

Measuring the similarity between symbolic sequences is essential in many data analysis. So far, due to the lack of natural geometrical interpretation of the symbolic sequence the resemblance (similarity) measure between two sequences was difficult to define [Kelil]. When spectra S1 and S2 of two sequences and are known, some natural similarity measures can be defined. In [13] the normalised number of a common set of mers

$$Si(C_1, C_2) = \frac{d\big(inter\,(S_1, S_2)\,\big)}{\sqrt{d\,(S_1)\,d\,(S_2)}} \tag{1}$$

was proposed as the similarity measure. Here $inter(S_1, S_2)$ is a set of mers that the two spectra share (intersection of S_1 and S_2), and $d\,(S)$ means the size (number of mers) of set S. The $Si(C_1, C_2)$ function varies between 0 when corresponding spectra are disjoint sets, and 1 when sequences C_1 and C_2 are identical (and their spectra as well).

The measure (1) has a geometrical representation in a multidimensional space spanned over mers. Any set of mers is represented by a binary vector whose components along mers belonging to the set are equal to one, while all others are equal to null. Now $d\,(S)$ is the length of the vector S and $d(inter(S_1, S_2))$ means the scalar product of the two vectors. The similarity given by (1) is simply the cosine of the angle between vectors S_1 and S_2.

The similarity measure that does not need any geometrical interpretation is also possible. It is the double fraction of mers that are common in both spectra against the total number of mers in the spectra

$$Si(C_1, C_2) = \frac{2d(inter(S_1, S_2))}{d\,(S_1) + d\,(S_2)}\,. \tag{2}$$

Both definitions yield comparable numbers when $d\,(S_1)$ and $d\,(S_2)$ differ by factor less than 2, otherwise (1) is significantly higher than (2). The last definition has the advantage that $1 - Si(C_1, C_2)$ can be considered as dissimilarity (not the distance) between S1 and S2 with the simple interpretation of being proportional to the fraction of the set theory the symmetric difference of S1 and S2 against the total number of mers in both spectra. In the following similarity definition (2) also known as the Dice coefficient will be used [14].

Representative set

The spectrum is a representation of a sequence in the space of relatively short distinct subsequences – mers. The important aim of a sequence analysis is to find proximity between a sequence and a set of sequences or between two sets of sequences. One widely used approach consists in computing all pair wise similarities between two sets followed by the assumption that an average similarity is the measure of their proximity.

A more appropriate method would be extracting the key features of a cluster to define a kind of fingerprint identifying the collection of sequences. We propose a set that includes mers common for spectra of a certain number of sequences (the union in the set theory) belonging to the cluster as the

cluster representative. It is a sort of a summary description of all objects contained in a cluster. We call the set the representative set (or spectrum) of the cluster. The representative set is significantly smaller than the original dataset. A similar method is frequently used in the analysis of categorical sequences [9,15,16], however, the representative set of categorical sequences consists of the smallest possible number (arbitrary to some extend) of sequences contained in a cluster.

In the present paper the representative set is precisely and uniquely defined. Let a cluster consist of n sequences $C_1, C_2, \ldots, C_n$, their corresponding spectra are $S1, S_2, \ldots, S_n$. The representative spectrum S_c of the set can be found in the following 3 steps. In the beginning $S_c(1) = S_i$, where S_i is spectrum of any sequence, next the spectrum S_j is selected and set of mers $diff(S_j, S_i)$ within spectrum S_j that are not in spectrum S_i is found. Sum $S_c(2) = S_i + diff(S_j, S_i)$ (union of S_i and S_j) becomes representative spectrum S_c of the two sequences. Continue $S_c(k) = S_c(k-1) + diff(S_c(k-1), S_{k-1})$ until $k = n$ (there is no spectrum left).

$$S_c = \lim_{k \to n} S_c(k) \tag{3}$$

In other words, $S_c(k)$ covering all sequences is the union of spectra $(S_1, S_2, \ldots, S_n)$.

Membership

The similarity between a particular sequence and a cluster set could be used as a measure of membership of some sequence in the given cluster. It seems more appropriate to define membership as the ratio of number of mers that the spectrum of the sequence shares with the representative spectrum of the cluster against the size of the spectrum

$$Ms(C) = \frac{d\left(inter\left(S, S_c\right)\right)}{d\left(S\right)}. \tag{4}$$

The membership varies from 0 when spectrum S and representative set of the cluster S_c are disjoint sets, to 1 when S is a subset of S_c.

III. EXAMPLE ANALYSIS

In this section the aforementioned notions are applied to clustering the mitochondrial genomes of four vertebrate groups: birds, fishes, mammals and turtles. The mitochondrial DNAs were downloaded from GenBank: http://www.ncbi.nih.gov/. The accession number of sequences can be found in supplementary material available on the web page:
http://cmst.eu/supp_mat/19_2/BKozarzewski.html .

Size of representative set

The representative set may be useful in (at least) two cases. First, if we are interested in selecting a rather small number of typical sequences that together summarize the main features of a whole set. We will show that the representative set covering a limited number of sequences is a useful tool in the clustering method. The representative set covers sequences $C_1, C_2, \ldots, C_n$ if membership of all their spectra are ones. Second, the representative set covering all sequences of interest is an efficient tool of lossless data reduction. In the present experiment the size (number of mers) of full coverage representative set $S_c(n)$ versus n (size of the data set) is analysed. The data set members are mitochondrial genomes of 200 mammal species. With the use of a parsing algorithm, spectra of sequences were obtained and the similarity matrix of sequences was calculated. The representative set covering one genome is about 2700 mers in size and grows to about 25824 mers for genomes of $n = 200$ species. It is approximately 5% of the total number of mers in spectra of all 200 genomes. Then the achieved reduction in data size is 95%.

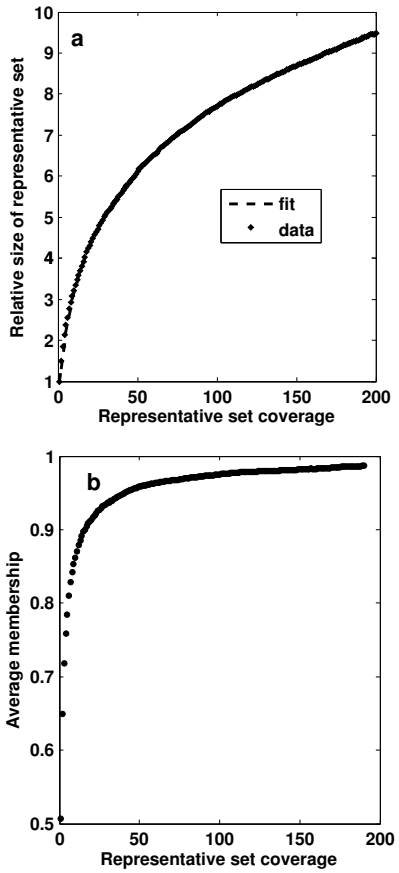

Fig. 1. a) Relative size of representative set, b) average membership of a single sequence

The function

$$S_c(n)/S_c(1) = 1 + 0.55(\log(n))^{1.64} \tag{5}$$

provides good fit to numerical results for the relative size of the representative set as shown in Fig.1a. Having the representative set covering k sequences out of 200, one can ask

about membership of the other 200-k mitochondrial DNA sequences.
From Fig.1b it follows that the average (over $200 - k$ sequences) membership reaches 0.90 when coverage is 15, and increases to 0.95 at coverage equal to 40. The representative set of 200 mitochondrial DNA sequences is a small and non-redundant subset (about 8800 mers) of the original dataset (about 540000 mers).

Hierarchical clustering (agglomerative)
In the beginning a set of 40 mitochondrial genomes including four groups (10 species each) is considered. The matrix consisting of similarities between each pair of spectra has been calculated. In the hierarchical clustering, the hierarchy of clusters is built starting with each spectrum as an individual cluster. The two most similar spectra are then grouped, giving one cluster of two species. The remaining clusters still consist of spectra of a single sequence. To find similarity between two clusters the Unweighted Pair Group Method Using Arithmetic Mean [17] is used. Within the cluster similarity is defined as arithmetic mean of similarities between all pairs of sequences that are members of the clusters. The clusters are then joined sequentially until all spectra are clustered. Unfortunately, there is no similarity-based method for stopping the clustering process, i.e. determining the correct number of clusters. In general, the best set of clusters is the one that keeps the distance between members in the same cluster small and the distance between members of adjacent clusters large. As a measure of distances the sum of squares within the clusters and sum of squares between all data are used. However, in our similarity-based method there is no natural distance measure available between sequences. However, a meaningful distance measure can be introduced, if rows of similarity matrix are considered as the vectors xi representing species in the procedure providing a reasonable stopping rule. Components of vector xi are similarities between sequence i-th and all sequences of the considered group. The Euclidean distance $|x_i - x_j|$ in the following as the measure between sequences i-th and j-th.

The three clustering criteria: a pseudo-F statistic [18], the Baesian information criterion [19] and the elbow point of percentage of the variance explained by a cluster versus number of clusters [20] will be used. The notations that will be used in the following are: N means size of data (number of sequences), N_c is number of clusters,

$$W = \sum_{i=1}^{N} |x_i - \bar{x}|^2$$

measures variability of all data,

$$W_k = \sum_{i=1}^{n_k} |x_i - \bar{x}_k|^2$$

measures intra-cluster variability of the k-th cluster, sum is over vectors belonging to the k-th cluster, n_k is size of k-th cluster, x_i is i-th row of similarity matrix representing i-th sequence, $\bar{x}$ and $\bar{x}_k$ are mean vectors of all sequences and sequences belonging to k-th cluster, respectively.

The pseudo-F statistic at a given step of clustering measures the variability of all data relative to the sum of variabilities within the clusters

$$pF = \frac{W - \sum_{k=1}^{N_c} W_k}{N_c - 1} \Big/ \frac{\sum_{k=1}^{N_c} W_k}{N - N_c} .$$

As clusters are joined, the pseudo-F statistic changes. A common recommendation on cluster selection is to choose a cluster size at which the values of the pseudo-F statistic are relatively high (compared to what is observed with other numbers of clusters). The Baesian information criterion is given by

$$BIC(N_c) = \sum_{k=1}^{N_c} n_k \log\left(\frac{n_k\,(n_k - N_c)}{W_k}\right)$$
$$-N\log(N) - 0.5[N - N_c^2 - N_c\log(N) - N^2\log(2\pi)],$$

and the percentage variance is defined as

$$PV(N_c) = \frac{100}{W}\sum_{k=1}^{N_c} W_k .$$

In the last two methods a knee (or elbow) point was determined by $F(n-1) + F(n+1) - 2F(n)$, where $F(n)$ is the index value and n is the current number of clusters. The first decisive local maximum is usually considered to be the correct number of clusters.

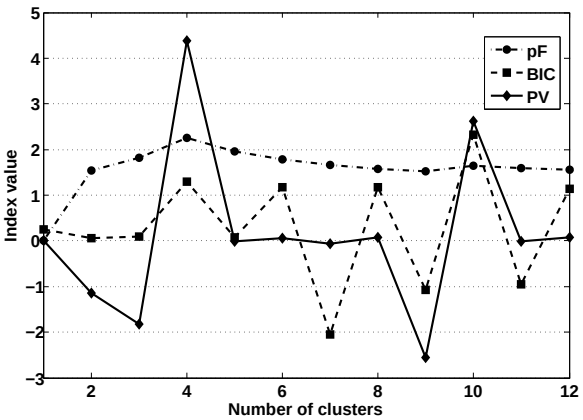

Fig. 2. Score of stop indices versus number of clusters

Figure 2 suggests four clusters as the best clustering result. At this level clustering yields: first cluster of 10 turtle species, second – 10 bird and 10 mammal species, third – 7 fish species and fourth – 3 fish species. Only 1 cluster is correctly and completely built. The clustering result is far from being satisfactory. Other clustering that could be considered results in 10 clusters, where all three indices also have a maximum. It seems more reasonable because 3 groups (bird, mammal

and turtle species) are correctly clustered but fish species are distributed between other 7 clusters. Nevertheless, this result is not satisfactory, either.

Hierarchical clustering based on representative sets
Here we propose an alternative clustering method based on determining small representative sets candidates covering several sequences (10 was found best) and then selecting sets of smallest variance. The selection uses some heuristic criteria for determining the representative set that at first builds a list of candidates for representative sets using a representativeness score and then eliminates redundancy [21].

In the present experiment a set of 100 genomes consisting of the four groups (birds, fishes, mammals and turtles), 25 species each, is considered. Spectra of sequences were obtained and the similarity matrix between all pairs was calculated. Candidates for representative sets are created as follows: at first the pair of sequences of largest similarity is selected from the similarity matrix, the union of their spectra becomes a seed of the candidate, the two sequences are removed from the pool of sequences. Then the similarity matrix is searched for the sequence of greatest membership in the seed. The union of the sequence spectrum and the seed becomes the new richer seed and the sequence is removed from the pool. The last step is continued until coverage of the seed reaches predefined number (10 in our experiment). In that way 10 candidates for representative sets were extracted. To define quality of the candidates the variance W_k/n_k was calculated for each candidate and a threshold (0.23 in our case) arbitrary to some extent is assumed. The candidates with variability exceeding the threshold are rejected. Therefore we were left with 8 candidate sets (covering 80 sequences) and 20 sequences to assign. In the next step the membership of every unassigned sequence in each of 8 candidate sets was calculated and the sequence was joined to the cluster of the highest membership score. Then the similarities between candidate sets were determined (Table 1), next the clusters were checked for potential redundancy. The two of very similar candidates are assumed to represent the same cluster.

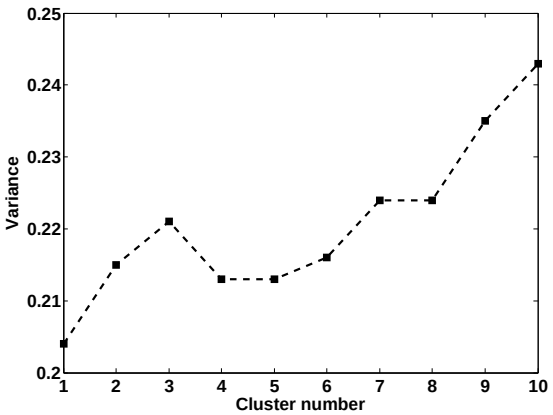

Fig. 3. Intra-cluster variance

The three (an arbitrary decision again) most similar pairs of candidate sets: (3;5) – similarity 0.68, (1;2) – similarity 0.67 and (4;8) – similarity 0.66 are assumed to belong to the three clusters.

Tab. 1. The similarity between pairs of candidates

pair of rep. set	similarity
(3,5)	0.677
(1,2)	0.674
(4,8)	0.659
(3,4)	0.655
(4,5)	0.655
(6,7)	0.653
(1,8)	0.652
(6,8)	0.651

After merging corresponding clusters we get 5 clusters, which we call B, F1, F2, M and T. Cluster B consists of 24 species, all are birds. Clusters F1 and F2 include 11 fish species and 12 species (11 fish, 1 mammal), respectively. Cluster M includes 28 species, 24 mammals, 1 bird and 3 fishes. Finally, cluster T includes 25 turtle species. One can say that in total 95% sequences were correctly assigned; however, only the turtle cluster is free of erroneous assignments. There is no indication that clusters F1 and F2 could be merged because most similar (with similarity 0.677) are M and T clusters as it follows from Table2 a). It might be interesting to compare the similarity matrices of the representative sets of the received clusters and the representative sets of the correct clusters, Table 2 b).

Tab. 2. The similarity matrix between: a) groups of species returned by the clustering algorithm and b) real groups of species

a)

	B	F1	F2	M	T
B		0.641	0.619	0.641	0.660
F1	0.641		0.622	0.637	0.639
F2	0.619	0.662		0.631	0.62
M	0.641	0.637	0.631		0.677
T	0.660	0.639	0.619	0.677	

b)

	B	F	M	T
B		0.675	0.672	0.661
F	0.675		0.681	0.665
M	0.672	0.681		0.679
F	0.661	0.665	0.679	

An unexpected observation is that most similar (0.681) are mitochondrial DNA representative sets of mammals and fishes. Hierarchical clustering based on representative sets provides much better results than agglomerative clustering but the user needs to choose an appropriate variance threshold to find the number of candidates for representative sets and a similarity threshold to merge most similar candidates.

Partitioning method based on representative sets
In partition-based clustering, the task is to partition a data set into a set of disjoint clusters of objects, so that each object is assigned to a unique cluster. Known representative sets of some sequences of known groups are used to categorize other sequences. The aim of the experiment is to determine the minimal size of representative sets that ensure a given percentage of correct assigning.

In the present paper the membership definition (4) is used as a measure for membership of the sequence in a representative set. The measure that is used as a criterion for membership of the object in a particular cluster is the highest membership score of the spectrum in the corresponding representative set. The data set to be partitioned consists of mitochondrial genomes of four groups (birds, fishes, mammals and turtles) each of 50 species. In the present approach four initial clusters, one from each of four groups of species is assumed and four representative sets are created.

The method generates the partitioned dataset as follows: the membership of each sequence in all four representative sets are calculated and a sequence is assigned to the cluster of highest membership. Representative sets remain unchanged and the resulting clusters are independent of the order in which sequences are processed. The method requires only one pass through the dataset. In the present experiment representative sets covering from 1 to 15 sequences randomly selected were created from each group and partitioning of all others sequences was performed. The algorithm was run 20 times to produce average scores presented in Fig. 4.

The clustering accuracy for birds and turtles is quite good even for small (coverage 5) RS while 95% accuracy for fishes is achieved only when coverage exceeds 15 and for mammals when coverage exceeds 25.

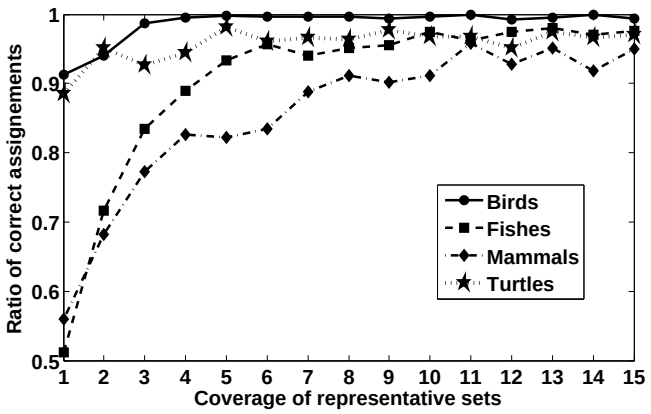

Fig. 4. Partitioning of the mitochondrial DNA into four groups of species

IV. CONCLUDING REMARKS

The problem of clustering large volumes of symbolic sequences has been considered. In our approach an original sequence of symbols over an alphabet is represented by a collection of short ordered and distinct subsequences of the symbols. The sequence analysis that followed was performed on a collection of subsequences called spectrum of the symbolic sequence. The discussed clustering algorithms use novel definitions of similarity between sequences and membership of a single sequence in a cluster of sequences, both are based on spectra. The performed experiments demonstrate that partitioning based on representative sets outperforms agglomerative (which is of the lowest performance) as well as hierarchical clustering based on representative sets (its performance is not bad but needs several arbitrary user decisions). One can find suggestions that moderate user interference into the clustering process can enhance the quality and performance of the clustering result. The result of our fourth experiment shows that if interference means independent generation of seeds of representative sets it can significantly help the processing of the representative-based clustering algorithm.

Acknowledgments

Author thanks an anonymous reviewer for his useful comments.

References

[1] M. Randić, S.C. Basak, *Characterization of DNA Primary Sequences Based on the Average Distances between Bases*, J. Chem. Inf. Comput. Sci. **41**, 561-568 (2001).
[2] Y. Liu, *The Numerical Characterization and Similarity Analysis of DNA Primary Sequences, Internet Electronic Journal of Molecular Design* **1**, 675-684 (2002).
[3] M-S. Yang and K-L. Wu, *A Similarity-Based Robust Clustering Method, IEEE Transactions on Pattern Analysis and Machine Intelligence* **2(4)**, 434-448 (2004).
[4] J. Wen, C. Li, *Similarity analysis of DNA sequences based on the LZ complexity*, Internet Electronic Journal of Molecular Design **6**, 1-12 (2007).
[5] A. Kelil, S. Wang, Q. Jiang, R. Brzezinski, *A general measure of similarity for categorical sequences*, Knowl. Inf. Syst. **24**, 197-220 (2010), (DOI 10.1007/s10115-009-0237-8).
[6] M.R. Ackermann, J. Blömer, D. Kuntze, C. Sohler, *Analysis of Agglomerative Clustering*, http://arXiv.org/abs/1012.3697 (2012).
[7] P. Berkhin, *Survey of Clustering Data Mining Techniques*,1-56, http://citeseerx.ist.psu.edu/viewauth/summary?aid=32145.
[8] R. Xu, D. Wunsch, *Survey of clustering algorithms*. IEEE Transactions on Neural Networks **16(3)**, 645-678 (2005).
[9] P. Agrawal, M.A. Alam, R. Biswas, *Analysing the agglomerative hierarchical clustering algorithm for categorical attributes*, International Journal Innovation, Management and Technology **1(2)**, 186-190 (2010) (and references quoted therein).
[10] N.S. Müller, A. Gabadinho, G. Ritschard, M. Studer, *Extracting knowledge from life courses: Clustering and visualization*, In DAWAK 2008, volume LNCS 5182 of Lectures Notes in Computer Science, 176-185, Berlin Heidelberg Springer (2008).

[11] G.W. Milligan, M.C. Cooper, *An examination of procedures for determining the number of clusters in a data set*, Psychometrika **50**, 159-179 (1985).
[12] D.-G. Ke , Q.-Y. Tong, *Easily adaptable complexity measure for finite time series*, Phys. Rev. E **77**, 066215 (2008).
[13] B. Kozarzewski, *A method for nucleotide sequence analysis*, CMST **18(1)**, 5-10 (2012).
[14] L.R. Dice, *Measures of the Amount of Ecologic Association Between Species*, Ecology **26(3)**, 297-302 (1945).
[15] M. Daszykowski, B. Walczak, D.L Massart, *Representative subset selection*, Analytica Chimica Acta **468(1)**, 91-103 (2002).
[16] A. Gabadinho, G. Ritschard, M. Studer, N.S. *Müller, Extracting and Rendering Representative Sequences*, in: *Communications in Computer and Information Science, Lecture Notes in Computer Science*, 94-106, Springer-Verlag Berlin Heidelberg (2011).
[17] C.D. Michener, R. R. Sokal, *A quantitative approach to a problem of classification*, Evolution **11**, 490-499 (1957).
[18] T. Calinski, J. Harabasz, *A Dendrite Method for Cluster Analysis*, Communications in Statistics **3(1)**, 1-27 (1974).
[19] Q. Zhao, V. Hautamaki, P. Fränti, *Knee point detection in BIC for detecting the number of clusters*, ACIVS 2008, volume LNCS 5295 of Lectures Notes in Computer Science, 664-673, Berlin Heidelberg. Springer (2008).
[20] V. Granville, *Identifying the number of clusters: final a solution*, http://www.analyticbridge.com/profile/Vincent.Granville
[21] M. Cameron, Y. Bernstein, H. Williams, *Clustered sequence representation for fast homology search*, J. Comp. Biol. **14(5)**, 594-614 (2007).

Bohdan Kozarzewski received PhD degree in physics (1965) at the Jagiellonian University in Cracow. Habilitated in 1975 in solid state theory. Since 1989 Professor at the Institute of Physics Technical University Cracow, sine 2006 at the University of Information Technology and Management, Rzeszow. Present research activity in computer modeling of nonlinear dynamical systems and time series analysis. Author and co-author of about 50 scientific publications.