ATLANTIS PRESS

Atlantis Studies in Uncertainty Modelling, volume 1

11th Conference of the European Society for Fuzzy Logic and Technology (EUSFLAT 2019)

A Hybrid Model of Gene Expression Profiles Reducing Based on the Complex Use of Fuzzy Inference System and Clustering Quality Criteria

Sergii Babichev[a] **and Jiří Barilla**[a,b] **and Jiří Fišer**[a,c] **and Jiří Škvor**[a,d]

[a]Department of Informatics, Faculty of Science, Jan Evangelista Purkyně University in Ústí nad Labem, 400 96, Ústí nad Labem, Czech Republic, sergii.babichev@ujep.cz

[b]jiri.barilla@ujep.cz, [c]jf@jf.cz, [d]jskvor@physics.ujep.cz

Abstract

The paper presents the hybrid model of gene expression profiles reducing based on complex use of fuzzy inference system and clustering quality criteria. The average of absolute values, the variance and Shannon entropy of the gene expression profiles were used as the input parameters of fuzzy inference system. The quality of gene expression profiles was used as output parameter of the model. The boundary value of the output parameter of the model was determined based on the minimum value of the clustering quality criteria which takes into account both the character of the objects distribution within clusters relative to the mass centers of the clusters, where these objects are, and the character of the clusters mass centers distribution in the feature space. Practical implementation of the proposed model allows us to divide the gene expression profiles into informative and non-informative objectively for purpose of the further investigation of the character of genes interconnection in the studied object.

Keywords: Gene expression profiles, Reducing, Fuzzy inference system, Clustering quality criteria, Shannon entropy, Statistical criteria.

1 Introduction

Gene expression profiles processing in order to study the character of genes interconnection in the studied objects is one of the current direction of modern bioinformatics. A solution of this problem promotes to create new effective antibiotics for treatment of complex diseases [12, 6]. Moreover, qualitative reconstructed gene regulatory network based on gene expression profiles will allow us to develop modern methods of diagnostic and treatment of complex diseases at genetic level. DNA-microarray technology [13, 16] and RNA-sequencing method [15, 17] are used to form the matrix of gene expression profiles nowadays. In any case we have as the result a high dimensional matrix, where number of rows is the number of the studied genes and number of columns is the number of the studied objects or conditions of the experiment performing. In paper [3] authors presented the results of the research concerning development of the information technology of gene expression profiles processing for purpose of gene regulatory networks reconstruction. The gene expression profiles reducing in terms of the statistical criteria and Shannon entropy is one of the stages of this technology implementation. The authors proposed to use the fuzzy inference system to divide the gene expression profiles into informative and non-informative in terms of the used criteria. The detail describe of this process implementation is presented in [1]. However, it should be noted that in these papers the authors do not justify the choice of the used criteria boundary values which are used for division of the gene expression profiles into informative and non-informative ones. The solution of this problem can be achieved with the use of modern methods of the data processing [18, 19, 14].

In this paper we propose the hybrid model of gene expression profiles reducing based on the complex use of both the fuzzy inference system and cluster analysis. Three groups of patients which were investigated on different types of Alzheimer disease were used during the simulation process. Detail description of the data is presented in the section "Materials and Methods". These data were divided into three clusters a priori. The simulation process involved the following stages: fistly, estimation of the gene expression profiles quality with the use of the statistical criteria and Shannon entropy as input parameters of the fuzzy inference system was performed. Then, the gene expression profiles were divided into informative and non-informative in

dependence on the quality criterion value. Finally, the clustering quality criterion was calculated for informative gene expression profiles. The optimal boundary values of the used criteria were determined based on the minimum value of the clustering quality criterion. The paper is organized as the following way: Section 2 presents the description of the investigated data and the hybrid model of the gene expression profiles reducing. This model is presented as the structural block chart of the algorithm of step-by-step process of the data processing implementation. This section contains also the clustering quality criterion which was used to determine the boundary values of the fuzzy inference system parameters. The results of the simulation concerning the investigated gene expression profiles reducing within the framework of the proposed model are presented in the section 3. This section contains also the discussion of the obtained results. The conclusions are presented in the section 4.

2 Materials and Methods

DNA microarrays data from database KEGG [7, 9, 8] were used as the experimental ones during the simulation process. This data contains 38 of DNA microarrays of patients which were investigated on Alzheimer disease [10, 11]. The DNA microarrays contain the information concerning genes expression of brain samples from three Alzheimer's Disease Centers. The first data contained 9 samples from entorhinal cortex (EC) of brain. The second data contained 10 samples from hippocampus (HIP) of brain. The third data contained 19 samples from primary visual cortex (VCX) of brain. Each of the samples contained 54675 of genes. So, the initial matrix of genes expression contained 38 of rows (samples) and 54675 of columns (genes). Gene expression profile in this case is presented as a vector of genes expression which are determined for different samples. The character of the genes expression values distribution in the investigated samples is presented in the figure 1. The analysis of the figure 1 allows us to conclude that the gene expression profiles can be divided into three distinguish clusters in dependence on type of the disease.

Three criteria were used for division of the gene expression profiles into informative and non-informative: variance, average of absolute values and Shannon entropy. The main idea for this process implementation is the following: if variance and average of the absolute values are less and Shannon entropy is larger than the appropriate boundary values, then this profile can be removed from the studied data as non-informative. In this case the gene expression values for various samples do not allow us to recognize the investigated samples. The Shannon entropy criterion was calculated based on James-Stein shrinkage estimator [5]. This method

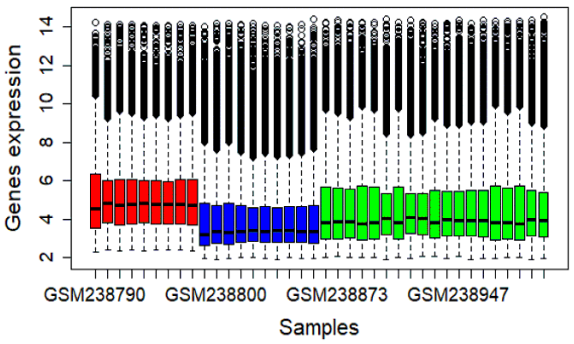

Figure 1: Boxplots of the investigated samples

is based on the complex use of two different models: a high-dimensional model with low bias and high variance, and a low-dimensional model with larger bias but smaller variance. The character of these criteria distribution in the studied gene expression profiles is presented in figure 2 and table 1.

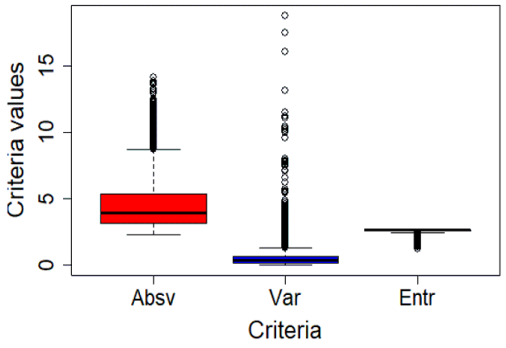

Figure 2: Boxplots of the statistical and Shannon entropy criteria distribution

Crit	Min	Quart 1	Med	Quart 3	Max
Abs	2.3	3.2	3.99	5.41	14.1
Var	0.01	0.2	0.39	0.64	18.8
Entr	1.22	2.61	2.67	2.7	2.71

Table 1: Statistical analysis of the used criteria distribution

Estimation of the gene expression profiles quality in terms of the used criteria was performed based on the fuzzy inference system with the use of Mamdani inference algorithm. The membership functions for both the input and output (quality of gene expression profiles) variables are presented in figure 3. As it can be seen, three linguistic terms for input variables (Low, Medium and High) and five for output

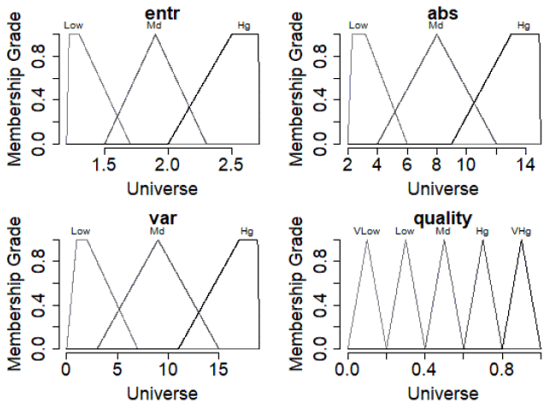

Figure 3: Membership functions of the fuzzy inference system

parameter (Very low, Low, Medium, High, Very high) were used within the framework of the proposed fuzzy inference system for purpose of the fuzzy rules formation. The range of the input variables change was divided into fifty equal sections during the simulation process. Implementation of the simulation process involves step-by-step increasing both the variance and average of the genes expression from minimum to maximum values and decreasing the Shannon entropy from maximum to minimum value. The value of the output parameter (quality) was estimated for each of the investigated gene expression profiles with the use of the fuzzy inference system. Conditions for division of the gene expression profiles into informative and non-informative are the following [1]:

$$var \leq var_{lim}; \quad abs \leq abs_{lim}; \quad entr \geq entr_{lim}; \quad (1)$$

If conditions (1) are true, the gene expression profile is removed from the database as non-informative. Otherwise, this profile is identified as informative and it is used for the following processing.
The next stage of the simulation process implementation involved division of the investigated objects which include only informative gene expression profiles into three clusters in accordance with the type of the data. The calculation of the clustering quality criterion was performed at this stage. The correlation distance was used as the metric to estimate the proximity level of the investigated vectors:

$$d(A,B) = 1 - \frac{\sum_{i=1}^{m}(x_{ai} - \bar{x}_a)(x_{bi} - \bar{x}_b)}{\sqrt{\sum_{i=1}^{m}(x_{ai} - \bar{x}_a)^2} \times \sqrt{\sum_{i=1}^{m}(x_{bi} - \bar{x}_b)^2}} \quad (2)$$

where m is the number of the informative genes; A and B are the studied objects; x_{ai} and x_{bi} are the expressions of the i-th gene for A and B objects respectively, $\bar{x}$ is the average value of genes expression of the appropriate vector.
The clustering quality criterion takes into account both the character of the objects distribution within the clusters and the character of the clusters mass centers distribution in the features space. The first component of this criterion is calculated as the average distance from the objects to the mass centers of the clusters where these objects are allocated:

$$QCW = \frac{1}{N}\sum_{s=1}^{K}\sum_{i=1}^{N_s} d(x_i^s, C_s) \quad (3)$$

where N is the number of the investigated objects; K is the number of the clusters; N_s is the number of the objects within the cluster S; C_s is the mass center of the cluster S; x_i^s is i-th object in the cluster S.
The second component of the clustering quality criterion is calculated as the average distance between the clusters mass centers:

$$QCB = \frac{2}{K(K-1)}\sum_{i=1}^{K-1}\sum_{j=i+1}^{K} d(C_i, C_j) \quad (4)$$

In [2] the authors proposed the complex internal clustering quality criterion which was calculated as multiplicative combination of both Calinski-Harabasz criterion [4] and WB-index [20]:

$$QC = \frac{K(K-1)QCW^2}{(N-K)QCB^2} \rightarrow min \quad (5)$$

This criterion was used during the simulation process. The structure block-chart of the algorithm for the investigated data processing within the framework of the proposed model is presented in figure 4. Its implementation involves the following steps:

1. Formation of the vectors of the fuzzy inference system input parameters: variance (var), average of gene expression profiles absolute values (abs) and Shannon entropy ($entr$). Setup of both the ranges and steps ($dvar$, $dabs$, $dentr$) of these parameters change.

2. Setup of the fuzzy inference system. Formation of the basic term set of the membership function for both the input and output variables and the set of fuzzy rules agreed between input and output parameters.

3. Initialization of the fuzzy inference system initial parameters: $var_1 = var_{min}$; $abs_1 = abs_{min}$; $entr_1 = entr_{max}$. Setup of the counter initial value of the fuzzy inference process implementation: $m = 1$.

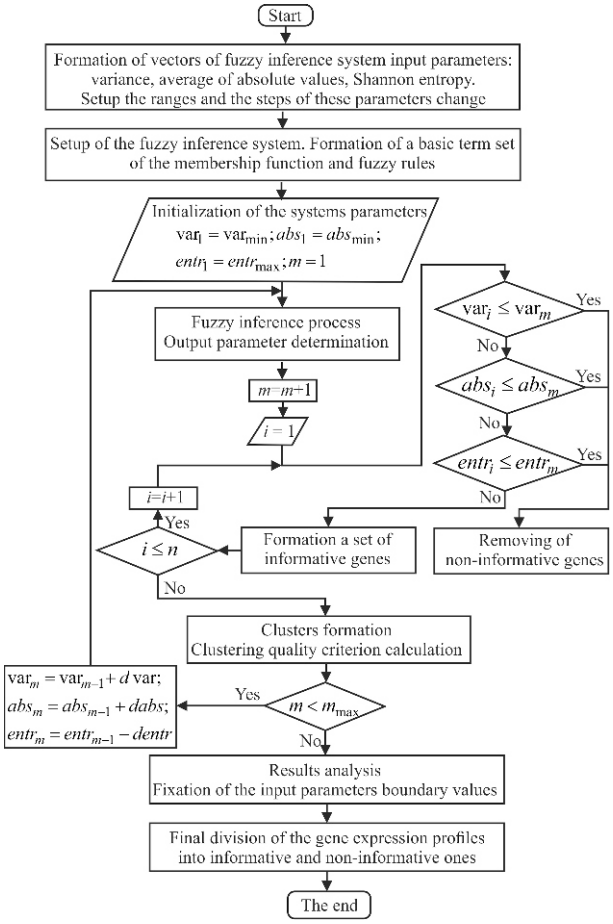

Figure 4: Structural block-chart of the algorithm for gene expression profiles reducing

4. Implementation of the fuzzy inference process for current values of the input parameters. Determination of the output parameter values (quality of gene expression profiles).

5. Division of the gene expression profiles into informative and non-informative taking into account the input parameters boundary values at appropriate stage of this process implementation according to the condition (1).

6. Formation of the clusters. Clustering quality criterion calculation with the use of the formulas (2)-(5).

7. If the counter value of fuzzy inference process implementation is less than maximum value, increasing of the boundary values of the input parameters and go to the step 4 of this procedure. Otherwise, results analysis and determination of the optimal boundary values of the input parameters which correspond to the minimum value of the clustering quality criterion.

8. Reducing of the gene expression profiles with the use of the optimal boundary values of the variance, Shannon entropy and average of the gene expression profiles absolute values.

3 Results and Discussion

Figure 5 presents the results of the fuzzy inference system simulation within the framework of the proposed model. As was described hereinbefore, the ranges of the input parameters were divided into fifty equal sections. The variance and the average of absolute values of the gene expression profiles were changed from minimum to maximum values (Fig. 5a,b) and the value of Shannon entropy was changed from maximum to minimum ones (Fig. 5c) during the simulation process. The value of the output parameter (quality of the gene expression profiles) was calculated at the each step of this process implementation (Fig. 5d). The charts of

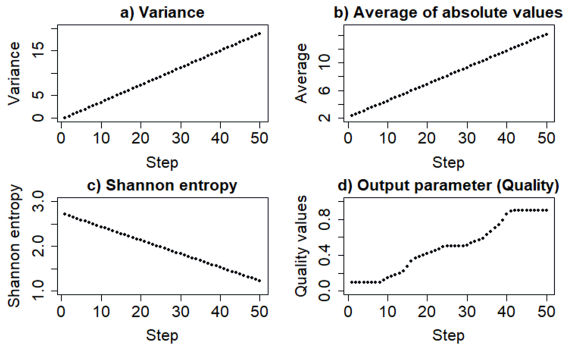

Figure 5: Results of the fuzzy inference system simulation

both the clustering quality criterion and the number of the informative gene expression profiles versus the step of the fuzzy inference process implementation are presented in figure 6. The analysis of the charts allows us to conclude that the number of the informative gene expression profiles is decreased monotonically during the change of the fuzzy inference system input parameters boundary values. At the same time, the clustering quality criterion value is changed chaotically. As it can be seen from Fig. 6a, the value of this criterion has three local minima during the simulation process. It means that the used gene expression profiles in these cases allow us to distinguish the investigated objects better in comparison with other cases. The number of the informative gene expression profiles in these cases are the following: 615 at 31-*st* step; 222 at 35-*th* step; 35 at 42-*nd* step. The quality of the gene expression profiles which were determined using fuzzy inference system belong to the range from medium to high values in the first and in the second cases. In

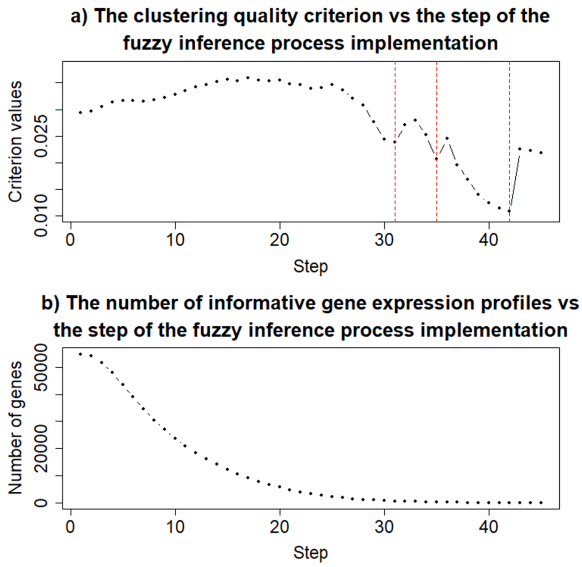

Figure 6: Charts of both the clustering quality criterion (a) and the number of the informative gene expression profiles (b) vs the step of the fuzzy inference process implementation

the third case (step = 42) the quality of the gene expression profiles was identified as very high one. This fact indicates the highest level of the genes expression profiles informativity in terms of separating ability of the investigated objects. The boundary values of the fuzzy inference system input parameters are the following: $var = 11.52$, $abs = 9.55$, $entr = 1.8$ in the first case ($step = 31$); $var = 13.05$, $abs = 10.31$, $entr = 1.679$ in the second case ($step = 35$); $var = 15.74$, $abs = 12.2$, $entr = 1.467$ in the third case ($step = 42$). The choice of the boundary values of the used parameters is determined by the aims of the current task. In the case of the further reconstruction of genes regulatory network the third case is the optimal one since the number of the allocated genes allows us to reconstruct the genes network qualitative for purpose of the following investigation of the character of genes interconnection taking into account the status of the object. In the case of the step-by-step gene expression profiles clustering and biclustering for purpose of genes regulatory networks reconstruction [3] the first or the second cases are the optimal ones since the number of the allocated gene expression profiles allows us to implement the gene expression profiles clustering and biclustering for both the reconstruction of genes regulatory networks and simulation of the genes networks obtained models.

4 Conclusions

In this paper we have proposed the hybrid model of gene expression profiles reducing based on the complex use of fuzzy inference system and clustering quality criterion. This model is presented as the algorithm of step-by-step data processing. The variance, the average of absolute values and Shannon entropy of gene expression profiles have been used as the boundary criteria for division of the gene expression profiles into informative and non-informative ones. Three groups of gene expression profiles of the patients which were investigated on Alzheimer disease have been used during the simulation process. The first data contained 9 samples from entorhinal cortex of brain. The second data contained 10 samples from hippocampus of brain. The third data contained 19 samples from primary visual cortex of brain. Each of the samples contained 54675 of genes.

The simulation process was performed in the following way. Firstly, the vectors of both the statistical criteria and Shannon entropy for the investigated gene expression profiles have been formed. The ranges of these criteria changes were divided into fifty equal sections and these values have been used for setup of the fuzzy inference system parameters. Then, the values of variance and average of absolute values of the gene expression profiles were changed monotonically from minimum to maximum and Shannon entropy from maximum to minimum values within the range of these parameters variation. The values of both the quality of gene expression profiles and the clustering quality criterion have been calculated at the each step of this process implementation. The optimal boundary values of the used criteria (var, abs, entr) were determined based on local minima of the clustering quality criterion. The results of the simulation have been shown that the best values of the used parameters in terms of the minimum value of the clustering quality criterion are the following: $var = 15.74$, $abs = 12.2$, $entr = 1.467$. In this case 35 of gene expression profiles were allocated. These genes can be used for both reconstruction of genes regulatory networks and simulation of the obtained models. However, it should be noted that the solution concerning determination of the parameters boundary values which correspond to the clustering quality criterion minimum values should be done taking into account the aims of the current task.

References

[1] S. Babichev, V. Lytvynenko, A. Gozhyj, M. Korobchynskyi, M. Voronenko, A fuzzy model for gene expression profiles reducing based on the complex use of statistical criteria and shannon en-

tropy, Advances in Intelligent Systems and Computing 754 (2019) 545–554.

[2] S. Babichev, V. Lytvynenko, M. Korobchynskyi, M. Taif, Objective clustering inductive technology of gene expression sequences features, Communications in Computer and Information Science 716 (2016) 359–372.

[3] S. Babichev, V. Lytvynenko, J. Skvor, M. Korobchynskyi, M. Voronenko, Information technology of gene expression profiles processing for purpose of gene regulatory networks reconstruction, in: Proceedings of the 2018 IEEE 2nd International Conference on Data Stream Mining and Processing, DSMP 2018, Lviv, Ukraine, 2018, pp. 336–341.

[4] T. Calinski, J. Harabasz, A dendrite method for cluster analysis, Communication in Statistics 3 (1974) 1–27.

[5] J. Hausser, K. Strimmer, Entropy inference and the james-stein estimator with application to nonlinear gene association networks, Journal of Machine Learning Research 10 (2009) 1469–1484.

[6] M. Čihák, Z. Kameník, K. Šmídová, N. Bergman, O. Benada, O. Kofronová, K. Petrícková, J. Bobek, Secondary metabolites produced during the germination of streptomyces coelicolor, Frontiers in Microbiology 8.

[7] M. Kanehisa, M. Furumichi, M. Tanabe, Y. Sato, K. Morishima, Kegg: new perspectives on genomes, pathways, diseases and drugs, Nucleic Acids Research 45 (2017) 353–361.

[8] M. Kanehisa, S. Goto, Kegg: Kyoto encyclopedia of genes and genomes, Nucleic Acids Research 28 (2000) 27–30.

[9] M. Kanehisa, Y. Sato, M. Kawashima, M. Furumichi, M. Tanabe, Kegg as a reference resource for gene and protein annotation, Nucleic Acids Research 44 (2016) 457–462.

[10] W. S. Liang, T. Dunckley, T. G. Beach, A. Grover, D. Mastroeni, D. G. Walker, R. J. Caselli, W. A. Kukull, D. McKeel, J. C. Morris, C. Hulette, D. Schmechel, G. E. Alexander, E. M. Reiman, J. Rogers, D. Stephan, Gene expression profiles in anatomically and functionally distinct regions of the normal aged human brain, Physiological Genomics 3 (28) (2007) 311–322.

[11] W. S. Liang, E. M. Reiman, J. Valla, T. Dunckley, T. G. Beach, A. Grover, T. L. Niedzielko, L. E. Schneider, D. Mastroeni, R. Caselli, W. Kukull, J. C. Morris, C. M. Hulette, D. Schmechel, J. Rogers, D. Stephan, Alzheimer's disease is associated with reduced expression of energy metabolism genes in posterior cingulate neurons, Proceedings of the National Academy of Sciences of the United States of America 11 (105) (2008) 4441–4446.

[12] K. Mikulík, J. Bobek, J. Zídková, J. Felsberg, 6s rna modulates growth and antibiotic production in streptomyces coelicolor, Applied Microbiology and Biotechnology 16 (98) (2014) 7185–7197.

[13] A. Prasad, S. M. A. Hasan, S. Grouchy, M. R. Gartia, Dna microarray analysis using a smartphone to detect the brca-1 gene, Analyst 1 (144) (2019) 197–205.

[14] Y. Rashkevych, D. Peleshko, O. Vynokurova, I. Izonin, N. Lotoshynska, Single-frame image super-resolution based on singular square matrix operator, in: Proceedings of the 2017 IEEE 1st Ukraine Conference on Electrical and Computer Engineering, UKRCON 2017, Lviv, Ukraine, 2017, pp. 944–948.

[15] W. Renthal, Localization of migraine susceptibility genes in human brain by single-cell rna sequencing, Cephalalgia 13 (38) (2018) 1976–1983.

[16] A. K. Shukla, P. Singh, M. Vardhan, Dna gene expression analysis on diffuse large b-cell lymphoma (dlbcl) based on filter selection method with supervised classification method, Advances in Intelligent Systems and Computing 711 (2019) 783–792.

[17] J. M. Tome, N. D. Tippens, J. T. Lis, Single-molecule nascent rna sequencing identifies regulatory domain architecture at promoters and enhancers, Nature Genetics 11 (50) (2018) 1533–1541.

[18] M. Štěpnička, S. Mandal, Fuzzy inference systems preserving moser–navara axioms, Fuzzy Sets and Systems 338 (2018) 97–116.

[19] L. A. Zadeh, A. M. Abbasov, S. N. Shahbazova, Fuzzy-based techniques in human-like processing of social network data, International Journal of Uncertainty, Fuzziness and Knowlege-Based Systems 23 (2015) 17.

[20] Q. Zhao, M. Xu, P. Fränti, Sum-of-squares based cluster validity index and significance analysis, in: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence) and Lecture Notes in Bioinformatics), Kuopio, Finland, 2009, pp. 313–322.