Acta Technica *62* (2017), No. 7A, 1535–1546

Research on targeted development of social value orientation and social autonomy based on semi-supervised clustering algorithm in the context of big data

Xinxin Liu[1]

Abstract. In order to improve the effectiveness of the research on targeted development of social value orientation and social autonomy in the context of big data, this paper proposes a research method of targeted development of social value orientation and social autonomy based on parallel clustering algorithm under the background of big data. Firstly, the intergenerational equity model is based to study the problem of targeted development of social value orientation, and the corresponding evaluation model is designed. Secondly, the semi-supervised clustering algorithm is introduced for analysis of the evaluation model for targeted development of social value orientation, which realizes the high-precision analysis of the targeted development of social value orientation. Finally, the simulation experiment is carried out to verify the effectiveness of the algorithm.

Key words. Big data, Semi-supervised clustering, Social value, Intergenerational equity.

1. Introduction

Value orientation is also values, which refers to the subjective evaluation paradigm used to measure the possibility of the value of things. It is reflected by people's behaviors and evaluation and attitude towards things. As the internal motivation that drives people's behaviors, it controls and regulates people's value behavior. People with different outlook on life and world outlook tend to have different values. With the continuous growth of a person, his/her values are constantly forming, and the formation of values is often influenced by factors such as family and society. In the

[1]School of Marxism Studies, Renmin University of China, Beijing, 100872, China

process of the formation of values, the social production mode and the economic status play a decisive role. Values can affect the behavior of a person, and it can also affect the behavior of the group and the behavior of the whole organization. Because people's values are different, they will have different behavior for the same thing. Every society has its common values of common recognition, which are the mainstream values related to the general trend of the development of the society.

With China's reform and opening up and rapid development, and communication with other countries in the world, some foreign culture has been introduced into China, leaving a profound impact on people, besides, the development of science and technology greatly enhances the popularity of the network, accelerating the speed of information dissemination, deepening the cultural exchanges, and contributing to the current multicultural development. Since the value orientation is influenced by the living environment and culture, the value orientation of pluralism has also developed naturally. In this increasingly diversified society, the pluralism of the value orientation is inevitable. Compared with the centralized society, the pluralistic society is obviously a manifestation of progress. Because centralization often represents a single thought, representing the rule of feudal centralization. In a centralized society, people's personality is often curbed, which will seriously hamper the rapid development of society. While pluralism represents individual freedom, and represents the real liberation of human nature, it is also the pluralism that gives rise to political democracy, in a pluralistic society, the free concepts of people are not contained, people can be more free to set their own direction, which is beneficial to the progress of science and technology, but also conducive to the rapid development of society. The diversity of values gives people more and more free choice and makes this society rich and colorful. However, while pluralism brings people benefits, it has also caused some social problems. Due to the rapid development of society and the deepening of pluralism, these social problems are also becoming increasingly serious. Therefore, precautions against these problems must be taken to make diversified values develop correctly.

Aiming at the targeted development of social value orientation and social autonomy in the context of big data, the intergenerational equity model is adopted in this paper to study the problem of targeted development of social value orientation, besides, the semi-supervised clustering algorithm is introduced to analyze the evaluation model of targeted development of social value orientation, and high-precision analysis on targeted development of social value orientation is achieved.

2. Intergenerational equity model of social value

The concept of intergenerational equity was first proposed by Talbot Page on the two bases of social choice and distribution justice, which mainly involves between the social welfare and resource allocation between contemporary and future generations. Page proposed the intergeneration view: the last generation provides a certain quantity and quality of wealth heritage to the next generation; the wealth heritage obtained by the next generation should be at least equal to the agricultural land resources inherited by the last generation. Intergenerational equity emphasizes

that the equitable distribution of agricultural land resources between present and future generations must be taken into account, the contemporary people should not only leave behind the agricultural land resources that satisfy their needs, but also ensure that the next generation can get the number and quality of agricultural land for their survival and development. The intergenerational equity of the distribution of agricultural land resources should include the following connotations:

The first connotation refers to the whether the distribution of agricultural land resources between generations is fair, which includes two meanings: first, whether the use of agricultural land resources by contemporary people destroys the basis for the development of future generations; second, whether contemporary people's investment in agricultural land resources matches the amount of agricultural land resources they consume. The investment in agricultural land resources includes the transformation of medium and low-yield fields, the construction of rural water conservancy infrastructure, land consolidation and so on.

The second connotation is that whether the contemporary people's compensation for future generations can be realized. Because the ideal state does not exist, if there is unfairness of cultivated land resources between present and future generations, and there are imbalance of cultivated land resources quantity and imbalance of function stock, contemporary people should make compensation for future generations. Where there is quantity imbalance, the level of science and technology can be improved to improve resource utilization efficiency and achieve intergenerational equity. Where there is absolute imbalance, active intergenerational wealth transfer must be subjected to for compensation for future generations.

Intergenerational wealth and social welfare equity are the physical presentation and level standard of intergenerational equity; therefore, intergenerational wealth and social welfare equity can ensure the realization of the intergenerational equity. In order to achieve the balance of intergenerational wealth, contemporary people must take the initiative wealth transfer policy, and establish special fund system of the intergenerational wealth transfer. A large number of facts prove that the depletion of resources, deterioration of ecological environment and the resulting economic and social problems have a warning behind which can not be ignored—compensation is not enough to realize the intergenerational wealth balance of agricultural land resources, with the establishment of intergenerational compensation mechanism of agricultural land, through the intergenerational transfer of agricultural land resources and the dynamic transfer between generations, reasonable compensation for future generations can be made, so as to achieve that in the practical application of intergenerational sharing and balanced development of natural resources, we can deal with intergenerational equity through the dynamic transfer of land resources between two generations. Specifically, the value of land resources is not reduced by intergenerational transfer, that is, the last generation not only transformed the physical quantity into the value quantity, but also made reasonable compensation for future generations through science and technology and other ways, so as to achieve the intergenerational sharing and balanced development of natural resources.

According to the above formula of$Y = X\,(1+i)^t$for intergenerational compensation, we need to find out the social compensation price for agricultural land of

contemporary people, $X$is the discount rate, $i$is the number of intergenerational years after which the intergenerational compensation price is calculated, Y is the social value of agricultural land and the social value of agricultural land based on the principle of intergenerational equity.

3. Semi-supervised clustering analysis

3.1. K-means algorithm

Semi-supervised clustering algorithm guides the clustering process by introducing a small amount of prior knowledge to the unsupervised clustering algorithm so as to improve the clustering performance. As a kind of classical unsupervised clustering algorithm proposed by MacQueen in 1967, the main idea of the K-means algorithm [42] is to assign each data object to the nearest class, that is, first to select K (K represents number of the cluster) objects as initial cluster center from N data object at random and then assign the rest of the data objects to the most similar (the shortest distance) class respectively according to the similarity (distance) between them and these clustering center. Every time iteration is completed, the mean of all the objects in each cluster is calculated as a new cluster center. The process is repeated until the clustering results begin to converge or the number of iterations can reach the maximum. The objective function of the K-means algorithm is defined as follows:

$$J_{K-means} = \sum_{j=1}^{k} \sum_{\boldsymbol{x}_i \in C_j} \|\boldsymbol{x}_i - \boldsymbol{u}_j\|^2 . \tag{1}$$

Where, $\boldsymbol{u}_j$ represents the central point of class C_j that the sample $\boldsymbol{x}_i$ belongs to, $J_{K-means}$ refers to the quadratic sum of the distance from data instances to centre of corresponding class. According to the target requirement of K-means algorithm, the smaller the $J_{K-means}$, the better it is. The algorithm process is shown in Algorithm 1.

Algorithm 1 K-means Clustering Algorithm Flow

Input: : number of cluster (K), data set X=$\{\boldsymbol{x}_0, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_{N-1}\}$;
Output: cluster classification C=$\{C_0, \ldots, C_{k-1}\}$;

1: Select K data instances as initial central point at random, such as C_0=$\boldsymbol{x}_1, \ldots, C_{k-1}$=$\boldsymbol{x}_{k-1}$;
2: Calculate the distance between $\boldsymbol{x}_0, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_{N-1}$ and central point $C_0, \ldots, C_{k-1}$ respectively. If the difference value with C_i is the minimum, the class is marked as i.
3: For all data instances marked as I, the mean is calculated as the new central point C_i;
4: Repeat steps 2 and 3 until the change of the Ci value is less than a given threshold or the number of iterations reaches the maximum

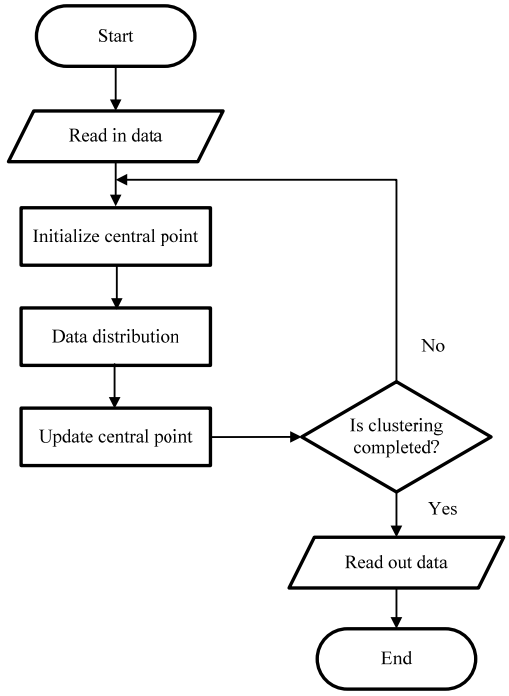

Fig. 1. K-means algorithm flowchart

3.2. Summary of semi-supervised clustering analysis

The semi-supervised learning has aroused wide attention of scholars since the 80s of the last century, which is a hot topic in the field of machine learning in recent years. Just as its name implies, semi-supervised learning is a kind of learning method which is between supervised learning and unsupervised learning. According to the combination of different mining methods and semi-supervised learning, it can be classified into three classes, that is, semi-supervised learning, semi-supervised clustering and semi-supervised regression. This paper mainly studies semi-supervised clustering which guides the clustering process by virtue of some known class labels or constraint information given, so as to improve the clustering performance. The semi-supervised also can be classified into two classes according to the different forms of introduced prior knowledge:

Semi-supervised clustering method based on data label and constraint guides the clustering process through known class label sample information to obtain more heuristic information so as to reduce the blind search. The common methods are as follows, i.e. 1. To satisfy the constraints compulsively, requiring that the result of clustering must satisfy the given constraint conditions.[9]. 2. Add penalty factor to objective function to punish the data instances which violates the constraint condition, so that the final clustering result can satisfy the constraint condition to the maximum, but the final clustering result may appear the situation of constraint

violation..[37]. 3. Constraint information is given in the form of separate class label information .[11, 38], also known as seed set (Seeds). The clustering center can be initialized with the help of these class labels so as to quicken the rate of convergence.

Distance-based semi-supervised clustering method: the characteristic of this algorithm is to find the distance measure function that satisfies the label or constraint by using the labeled data object and then adjust the distance between samples by this measure function, requiring that the adjusted sample distance must satisfy a given constraint conditions. The common methods are as follows, i.e. 1. Obtained mahalanobis distance based on convex optimization algorithm [39]. 2. Obtained Euclidean distance based on the shortest distance. 3. Obtained Jensen-Shannon discrete magnitude by gradient descent [40]. 4. Improve the editing distance [41] (string-edit distance) by EM algorithm.

3.3. Form of prior knowledge

Prior knowledge, also known as background knowledge or domain knowledge, is a set of specific constraint information. The prior knowledge is introduced into the traditional unsupervised clustering algorithm to guide the clustering process, which can improve the clustering quality. The representation of prior knowledge depends on the specific application field. The most commonly used prior knowledge in the semi-supervised algorithm is as follows:

(1) Constraint relation between positive association and negative association

A constraint relation, also known as a constraint condition, is a constraint relation to a data object, and its principal is two data instances. The positive correlation constraint relation indicates that two data instances are of the same class, expressed in *Must-Link*. The negative association constraint relation indicates that two data instances belong to different classes, expressed in *Cannot-Link*. As it can describe the relationship between two data instances, it is often referred to as pairwise constraint.

Assuming that two data instances, namely *xi* and *xj* , belong to class *Ci* and *Cj* respectively, and if (*xi*,*xj*)∈*Must-Link*, then $i = j$. If (*xi*,*xj*)∈*Cannot-Link*, then $i \neq j$. The constraint of *Must-Link* and *Cannot-Link* has symmetry and transitivity:

a) Symmetry

$$(x_i, x_j) \in Must\text{-}Link \Rightarrow (x_j, x_i) \in Must\text{-}Link\,. \tag{2}$$

$$(x_i, x_j) \in Cannot\text{-}Link \Rightarrow (x_j, x_i) \in Cannot\text{-}Link\,. \tag{3}$$

b) Transitivity

$$\begin{gathered}(x_i, x_j) \in Must\text{-}Link \&\& (x_j, x_k) \in Must\text{-}Link \\ \Rightarrow (x_i, x_k) \in Must\text{-}Link\,.\end{gathered} \tag{4}$$

$$\begin{gathered}(x_i, x_j) \in Must\text{-}Link \&\& (x_j, x_k) \in Cannot\text{-}Link \\ \Rightarrow (x_i, x_k) \in Cannot\text{-}Link\,.\end{gathered} \tag{5}$$

(2) Partial class label information

All class label information in the unsupervised algorithm is unknown. The label information of some data objects can be introduced to the clustering process by these sample information once obtained, so as to guide the clustering process. Assuming that the dataset object to be analyzed is D, and partial dataset object of known class label is P, and dataset object of unknown class label is U, then there is P∪U=D, P∩U=∅. If data instance xi∈P, then xi is able to get the label information of its class through the relevant knowledge. If xi∈U, then the class label information of xi is unknown. Under normal conditions, P<<U, that is, the data of an unknown class label is much larger than that of a known class label.

(3) Rule information

Rule information refers to the association relationship among properties of data objects excavated through related algorithm of association rules, and these rules can be used to guess the unknown attribute values of data objects. For UCI dataset *Weather,* the following association rule information can be obtained by *Aporior* algorithm:

outlook=rain&temperature=cold =>play=no

The associated information between *outlook*, *temperature* and *play* can be obtained through the above formula. If *outlook=rain* and *temperature=cold*, then the probability of *play=no* is 1.

3.4. Cop-Kmeans algorithm

Cop-Kmeans algorithm [9] is more commonly used semi-supervised clustering algorithm which introduces the pair-wise constraint information to the K-means algorithm, and its basic clustering idea is the same as that of K-means. However, it is required that the data object must satisfy constraint condition of *Must-Link* and *Cannot-Link* during the data distribution process. Assuming that the number of cluster for some dataset K=2, $C1$ and $C2$ represent two classifications of such dataset respectively, $u1$ and $u2$ represent the central point of each classification respectively, and $\boldsymbol{x}i$ and $\boldsymbol{x}j$ are two data instances. The full line represents a *Must-Link* constraint between two data instances, while the imaginary line represents a *Cannot-Link* constraint.

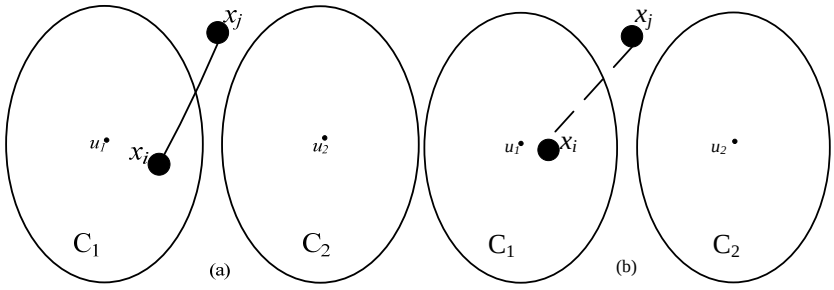

Fig. 2. Two constraints of cop-kmeans algorithm during data distribution

As shown in Fig.2 (a), if $\boldsymbol{x}i$ has been distributed to the closer class $C1$ by calculating distance, and $\boldsymbol{x}j$ is the current sample to be distributed, Cop-Kmeans algorithm

will directly distribute $\boldsymbol{x}j$ to class $C1$ instead of calculating the distance between sample $\boldsymbol{x}j$ and central point of two classes due to the constrained relationship of positive association between $\boldsymbol{x}i$ and $\boldsymbol{x}j$, namely, $(\boldsymbol{x}i, \boldsymbol{x}j) \in$*Must-Link*. Even if $\boldsymbol{x}j$ is likely to be much closer to central point of class C2, two data instances must be assigned to the same class according to the *Must-Link* constraint requirements.

As shown in Fig.2 (b), if $\boldsymbol{x}i$ has been distributed to the closer class $C1$ by calculating distance, and $\boldsymbol{x}j$ is the current sample to be distributed, Cop-Kmeans algorithm will directly distribute $\boldsymbol{x}j$ to class $C2$ instead of calculating the distance between sample $\boldsymbol{x}j$ and central point of two classes due to the constrained relationship of negative association between $\boldsymbol{x}i$ and $\boldsymbol{x}j$, namely, $(\boldsymbol{x}i, \boldsymbol{x}j) \in$*Cannot-Link*, to ensure that it is distributed to different class with $\boldsymbol{x}i$. If the number of cluster is greater than 2, $\boldsymbol{x}j$ will be distributed to the closer class except for class $C1$. Based on above analysis, the Cop-Kmeans algorithm flow can be described as follows.

Algorithm 2 Cop-Kmeans Clustering Algorithm Flow

Input: : number of cluster K, dataset $X=\{\boldsymbol{x}0, \boldsymbol{x}1, \ldots, \boldsymbol{x}N-1\}$;
Output: cluster classification $C=\{C1, \ldots, Ck\}$;

1: Select K data instances as initial central point at random, such as $C0=\boldsymbol{x}1, \ldots, C\boldsymbol{k}\text{-}1=\boldsymbol{xk}\text{-}1$;
2: Select partial data from the dataset at random to generate *Must-Link* constraint set *ML* and *Cannot-Link* constraint set *CL* respectively;
3: For any sample to be distributed $\boldsymbol{x}i$, if $(\boldsymbol{x}i, \boldsymbol{x}j) \in$*Must-Link* and $\boldsymbol{x}j$ is distributed to class *Ci*, then $\boldsymbol{x}j$ will also be distributed to the class *Ci*. If $(\boldsymbol{x}i, \boldsymbol{x}j) \in$*Cannot-Link* and $\boldsymbol{x}j$ is distributed to class *Ci*, then $\boldsymbol{x}j$ will be distributed to the closer class *Cj* in addition to *Ci*. Otherwise, $\boldsymbol{x}i$ will be distributed to the closer class *Ci*;
4: Update the central point of each class once the iteration is completed.
5. Repeat step 3 and 4 until the change of *Ci* is less than given threshold or the number of iterations reaches the maximum.

4. Experimental analysis

Targeted at social value assessment of land resources under the principle of intergenerational equity, this experiment analyzes and verifies the assessment performance of the above algorithm. On the basis of current social value X of farmland, it is necessary to select a certain land reduction rate i and a certain intergeneration term t to obtain the social security price Y of the farmland under the principle of intergenerational equity.

Land reduction rate is the interest rate that restores the pure land income to the land price, of which the essence is a rate of return on capital investment. In general, the size of the rate of return is positively related to the size of the investment risk, that is, the greater the risk is, the higher the rate of return will be. On the contrary, the smaller the risk, the lower the rate of return is. Determining the appropriate

reduction rate is one of the key problems of accurate calculation of land price. At present, the commonly used methods to determine the rate of land reduction in China are generally as follows. (1) Replaced with benchmark one-year deposit rate or mean of bank deposit rate for many years. (2) Determination of the proportion of land rent and land price in the past year. (3) The mean of bank deposit rate for many years is added and subtracted or divided by the price index. (4) Use annual rate of investment of local ordinary bank as reduction rate. (5) The reduction rate is obtained by the method of safety interest rate plus risk adjustment value. (6) Real interest rate means the interest rate after adjustment by price index and deducting 10% income tax based on benchmark one-year deposit rate. The above 6 kinds of calculation methods for land reduction rate are evaluated from rationality and the difficulty and easy of determining rate. The first is the simplest and appropriate, but the reasonableness is poor, because the bank rate has been adjusted several times within one year in recent years, and the price is more stable. The different rate of the same farm land use is evaluated in current year. It is obviously not very reasonable as the land price is very different. The second method is more scientific, but the current rural land property market is still in the gestation, and the data collection is difficult. The fourth method also has difficulties in collecting data because there is less research on investment interest rates on agriculture. The third, fifth, and sixth method are widely used currently. The social security price of the male population land expropriation area is shown in Table 1.

Table 1. Social security price of male population land expropriation area

Age group	Amount (yuan / person)	Population ratio (%)	Insurance base (yuan / person)	Insurance standards (yuan / person)	Education fee (yuan / person)	Training fee (yuan / person)	Total number (person)
0∼18	7600	17.31	51	120	3000	–	12.366
18∼40	12000	22.07	83	120	5000	–	15.768
40∼60	19800	14.23	168	120	–	2000	10.163
> 60	25000	7.35	–	120	–	–	5.231

Table 2. Comparison of running time and convergence precision

Test data selection	Index	[2]Document [2]	[3]Document [3]	[5]Document [5]	Algorithm in this paper
Situation 1	Convergence precision	2.68	0.35	5.35	2.32E-3
	Calculation time	4.57	9.45	4.42	3.21
Situation 2	Convergence precision	3.42	1.18	6.09	3.57E-2
	Calculation time	4.96	9.91	4.73	3.53

The quantization comparison of convergence time and precision index of selected centralized algorithm is shown in Table 2. Among the two indexes mentioned above, the improved algorithm is optimal. The convergence precision of this algorithm is better. However, it is relatively better to calculate the time because of parameter optimization is added, convenient for practical application.

5. Conclusion

Every society is bound to have its mainstream value that reflects the essential requirement of the social system and basic spirit of this society although today's society is in value diversification. If people's values deviate, it will lead to social confusion and mental impetuosity, especially in this value diversification. At the same time, it is easier for people to get lost in many values in the period of social and economic transformation. Therefore, it is essential to guide people to correct values. We can start from the following points, that is, the first is to guide people to be a rational person and to look at the value diversification rationally. If people are very rational, they will not be lost in many values. Instead, they can see clearly the right and mainstream values to avoid social impetuosity caused by the loss of values. The second is to propagandize socialist core values actively and coordinate with education of advanced culture in the meantime. Only in this way can a good effect be achieved. The third is to emphasize personal value in addition to the importance of the social value in the guidance of people's values. The value of the individual is constantly highlighted in the pluralistic society. People can observe the moral of the society more consciously if guidance is carried out correctly.

References

[1] Y. Y. Zhang, A. Algburi, N. Wang, V. Kholodovych, D. O. Oh, M. Chikindas, and K. E.. Uhrich: *Self-assembled Cationic Amphiphiles as Antimicrobial Peptides Mimics: Role of Hydrophobicity, Linkage Type, and Assembly State*, Nanomedicine: Nanotechnology, Biology and Medicine *13* (2017), No. 2, 343–352.

[2] Y. Song, N. Li, J. Gu, S. Fu, Z. Peng, C. Zhao, Y. Zhang, X. Li, Z. Wang, X. Li: *β-Hydroxybutyrate induces bovine hepatocyte apoptosis via an ROS-p38 signaling pathway.* Journal of Dairy Science *99* (2016), No. 11, 9184–9198.

[3] N. Arunkumar, K. R. Kumar, V. Venkataraman:*Automatic detection of epileptic seizures using new entropy measures.* Journal of Medical Imaging and Health Informatics *6* (2016), No. 3, 724–730.

[4] R. Hamza, K. Muhammad, N. Arunkumar, G. R. González: *Hash based Encryption for Keyframes of Diagnostic Hysteroscopy*, IEEE Access, https://doi.org/10.1109/ACCESS.2017.2762405 (2017).

[5] J. W. Chan, Y. Y. Zhang, and K. E. Uhrich:Amphiphilic Macromolecule Self-Assembled Monolayers Suppress Smooth Muscle Cell Proliferation, Bioconjugate Chemistry *26* (2015), No. 7, 1359–1369.

[6] Y. J. Zhao, L. Wang, H. J. Wang, and C. J. Liu: *Minimum Rate Sampling and Spectrum Blind Reconstruction in Random Equivalent Sampling.* Circuits Systems and Signal Processing *34* (2015), No. 8, 2667–2680.

[7] S. L. Fernandes, V. P. Gurupur, N. R. Sunder, N. Arunkumar, S. Kadry:*A*

novel nonintrusive decision support approach for heart rate measurement. Pattern Recognition Letters. https://doi.org/10.1016/j.patrec.2017.07.002 (2017).

[8] N. Arunkumar, K. Ramkumar, V. Venkatraman, E. Abdulhay, S. L. Fernandes, S. Kadry, S. & Segal:*Classification of focal and nonfocal EEG using entropies.* Pattern Recognition Letters *94* (2017), 112–117.

[9] J. W. Chan, Y. Y. Zhang, and K. E. Uhrich:*Amphiphilic Macromolecule Self-Assembled Monolayers Suppress Smooth Muscle Cell Proliferation*, Bioconjugate Chemistry *26* (2015), No. 7, 1359–1369.

[10] M. P. Malarkodi, N. Arunkumar, N., V. Venkataraman: *Gabor wavelet based approach for face recognition.* International Journal of Applied Engineering Research *8* (2013), No. 15, 1831–1840.

[11] L. R. Stephygraph, N. Arunkumar: *Brain-actuated wireless mobile robot control through an adaptive human-machine interface.* Advances in Intelligent Systems and Computing *397* (2016), 537–549.

[12] D. S. Abdelhamid, Y. Y. Zhang, D. R. Lewis, P. V. Moghe, W. J. Welsh, and K. E. Uhrich:*Tartaric Acid-based Amphiphilic Macromolecules with Ether Linkages Exhibit Enhanced Repression of Oxidized Low Density Lipoprotein Uptake*, Biomaterials *53* (2015), 32–39.

[13] W. Pan, S. Z. Chen, Z. Y. Feng: *Automatic Clustering of Social Tag using Community Detection.* Applied Mathematics & Information Sciences *7* (2013), No. 2, 675–681.

[14] Y. Y. Zhang, E. Mintzer, and K. E. Uhrich: (2016) *Synthesis and Characterization of PEGylated Bolaamphiphiles with Enhanced Retention in Liposomes*, Journal of Colloid and Interface Science *482* (2016), 19–26.

Received May 7, 2017