Innov Biosyst Bioeng, 2018, vol. 2, no. 2, 84–89
doi: 10.20535/ibb.2018.2.2.133466
UDC 004.021

ESTIMATION OF ALGORITHMS EFFICIENCY IN THE TASK OF BIOLOGICAL OBJECTS CLUSTERING

V.S. Umanets*, B.A. Voinyk, V.A. Pavlov, Ie.A. Nastenko

Igor Sikorsky Kyiv Polytechnic Institute, Kiyv, Ukraine

*Corresponding author: 2_bytes@ukr.net

Received 3 June 2018; Accepted 26 June 2018

Background. The task of determining the functional relationship between biophysical parameters is an integral part of the actual problem of finding the optimal impact on a biological object and is currently not completely resolved. One of the important tasks in this area is the partitioning of the original feature space into such areas (clusters) that relate to different functional relationships linking biophysical parameters and have, in general, an arbitrary shape. Such clusters in the future is logical to call functional. To obtain and analyze the functional clusters, there are a number of algorithms, each of which has its advantages and disadvantages. At the same time, the solution of a certain practical problem requires an evaluation of the efficiency of the algorithms in terms of the cluster separation adequacy.

Objective. In this paper, for a general example of the biological objects clustering problem (Fischer's Iris Data Set), the efficiency of a typical clustering tools series is evaluated. The application of k-means classical algorithm, the Ward algorithm and developed in this work the fuzzy version of clustering for the k-means algorithm with a limited mass of the working area for the clusters' formation was considered.

Methods. The algorithm includes a procedure for a priori estimation of the clusters quantity. The estimation is carried out according to the frequency histogram. To determine the optimal number of the histogram columns, the application of the Scott formula is justified. The algorithm allows forming clusters of arbitrary configuration with obtaining the value of the object's membership measure for each of the clusters. The comparative testing of the above algorithms was carried out on Fisher's Iris Data Set.

Results. The best value of F_1-score is obtained for the algorithm proposed in this paper: $F_1 = 0.92$, the value $F_1 = 0.90$ is obtained for the Ward method and the value $F_1 = 0.88$ – for the classical *k*-means algorithm.

Conclusions. The obtained test results on the analysis problem of arbitrary-shaped clusters made it possible to give preference to the version of fuzzy *k*-means with a limited mass of the working area for the clusters' formation. The calculating of the membership measure value allows us to obtain additional information on the structure of cluster formations, as well as to correct the result of clustering of k-means with a limited mass, which is especially important since the formation of clusters occurs in a single pass. Comparing the computational resources required for computing algorithms with relatively close test results also makes it possible to give preference to the developed algorithm. Compared with the Ward algorithm, it requires fewer computing resources since no additional memory is needed to store the distance matrix and no time is required to recalculate it.

Keywords: clustering; *k*-means; biological object; membership function; estimation of a number of clusters; fuzzy clustering.

Introduction

The task of determining the functional relationship between biophysical parameters is an integral part of the actual task of finding the optimal impact s on biological objects and is not fully resolved now. Among the most interesting results of such tasks are the ones that adequately represent groups in initial data with different functional relationships that bind considered biophysical parameters. It is logical to call such clusters functional, and their form in the general case can be arbitrary. To obtain and analyze the functional clusters, there are a number of algorithms, each of which has its advantages and disadvantages. At the same time, the solution of a certain practical problem requires an evaluation of the efficiency of the algorithms in terms of the cluster separation adequacy. Such an evaluation of algorithms can be performed on a given set of objects, the classification of which is known to the researcher but is unknown to the tested algorithms.

One of the most common approaches to clustering of multidimensional data are the methods of *k*-means family [1]. However, the classical version of the approach tends to form exclusively multidimensional spherical clusters by minimizing each cluster's dispersion. One of the ways to overcome this problem is introduction of a limitation on the total mass of the working area, by which the current value of the centroid of the cluster is determined. One of the current versions of the algorithm [2] implements this approach. However, this version of the algorithm has a number of shortcomings: the need to set the number of groups before the clustering and the lack of a mechanism for calculating the membership degree to the cluster. Note that a number of methods based on information entropy [3, 4] and divergence [3, 5] solve the first of these problems, but they are quite loaded from a computational point of view, so it is desirable to have a simpler mechanism for obtaining this estimate. Below we propose to develop an algorithm for fuzzy *k*-means with a limited mass of the cluster formation working area. Then evaluate the effectiveness of the developed algorithm comparing to the classical algorithm of k-means and one of the most frequently used hierarchical algorithms – Ward's method.

The work's purpose is to develop a version of the *k*-means method, which solves the problem of partitioning the initial sample by forming clusters of arbitrary form.

The study objectives are to develop a version of the fuzzy clustering algorithm for the *k*-means method with a limited mass of the working area of cluster formation, the introduction into the algorithm a simple mechanism for a priori estimation of the clusters quantity and to evaluate the algorithm efficiency on a biological objects sample with known cluster partitioning.

Materials and methods

The approach justification. The *k*-means algorithm standard mechanism without limiting the mass (quantity) of the working area objects of cluster formation leads to forming the spherical shape clusters which are identified as the ideal form of object groups. In this case, the centroids' path to the limiting state is neither an object of the analysis of the algorithm nor a constructive element used in the formation of the cluster. Only the stability of the centroid boundary state is important, which determines the result of clustering.

In [2], the mechanism of the k-means algorithm was first used to obtain non-spherical clusters, while the basis for determining the form of the received cluster is no longer the boundary position of the centroid, but the path at which the working area centroid passes into its limiting state. The displacement of the centroid determines the trend of the cluster working area and actually allows the algorithm to recognize its fragments. However, with the implementation of the standard *k*-means mechanism as the new objects are joined to the working area, the centroid travel speed steadily decreases. This happens due to reduction in the weight of the attached object impact in relation to the previously accumulated mass of the working area and therefore new object's impact on the trend becomes insignificant. The introduction of the working area with limited mass of cluster formation in [2] allowed to propose a mechanism forming of arbitrary form clusters and to extend the method of *k*-means to the task of cluster analysis general case. However, as noted above, this algorithm version may be expediently supplemented by estimating the clusters quantity in this sample of data and by calculating the objects' measure of membership to clusters.

Clustering algorithm. Let object x_j, $j = 1,...,n$ is described as a string j $\{x_{j1}, x_{j2},...,x_{jm}\}$ of initial matrix M of dimension $\dim(n,m)$.

The algorithm implements the following steps:

1. Normalizing data.

2. Initial centroid initialization using one of several methods:

(a) close to the zero vector;

(b) close to the vector center of mass;

(c) initialization on the peripheral data points;

(d) centroids are located evenly distant from the center of mass with a given step;

(e) the position of the centroids is chosen randomly

3. Object x_l is chosen and distances between x_l and every of k_t centroids are being calculated;

4. The object is joined to the nearest cluster.

5. The centroid moves to the new position calculated using following formulas:

if $n_t < I_{\max}$, then

$$x_{c_t} = \frac{\sum_{i=1}^{n_t} x_i + x_l}{n_t}$$

if $n_t < I_{\max}$, then

$$x_{c_t} = \frac{\sum_{i=1}^{n_t} x_i + x_l - p \cdot o}{I_{\max} - p}, \; n_t = n_t - p \qquad (1)$$

where n_t is a number of points in cluster t, which are used to calculate a centroid new position, $I_{\max}$ is the point's number threshold, C_t is the name of cluster t, sum $\sum_{i=1}^{n_t} x_i + x_l$ accumulates information or calculating the centroid location, p is the number of conditional objects *o* whose coordinates are the coordinates of current centroid's position.

As can be seen from formula (1), if the threshold of points' number is reached, "part" of the pre-accumulated information is "forgotten", which allows controlling movement of the centroid in the process of clustering. An adequate choice of parameters provides more ordered motion when reproducing the functional dependence.

6. The procedure stops after processing all of N points unless different stopping condition specified.

Clustering is carried out in a one-pass, and clusters obtained as a result have a non-spherical form.

The algorithm described above was proposed in [2] however, it needs the number of clusters to give an adequate result. The mechanisms of a clusters quantity priori estimation are rare. For the most part data scientist use posterior methods like gradually lowering the number of clusters together with using quality estimation methods. To address this problem, we propose the next approach. A density distribution histogram is constructed for each of the *m* variables. Using the number of local maxima in each variable's distribution, one can obtain an estimate of clusters quantity in the sample. The problem of columns optimal number can be solved by using the Scott formula, the Friedman–Diakonis formula, or similar ones. In the implementation of the algorithm, the Scott formula [6] was used, due to lower computational cost compared to the Friedman–Diakonis formula that uses interquartile distance thus requiring data to be ranked.

It was proposed to add membership function calculation to the algorithm. However, to use the known approach for this purpose, similarly to the *C*-means algorithm [7], it is incorrect, since the calculation of the centroid new position in *C*-means occurs after the information accumulation and not in the centroids' movement process. In the case when the position of the centroid changes in the process of adding points, calculated value of the membership function will lose its relevance. In this case, the value of the membership function should be calculated already after the initial formation of clusters. In addition, the mechanism of membership degree calculating also requires to be changed, as in contrast to the classical version of the *C*-means (the formation of clusters of hyperspherical form), the clusters received will in most cases have a stripe-like form

To solve the problem, two possible ways can be suggested:

1. Evaluate membership function using average distance between an object and other objects in the cluster can be used:

$$u_{ij} = \frac{1}{\sum_{k=1}^{c}\left(\dfrac{\dfrac{1}{n_j - 1}\sum_{l=1}^{n_j}\|x_i - x_l^{(j)}\|}{\dfrac{1}{n_k}\sum_{l=1}^{n_k}\|x_i - x_l^{(k)}\|}\right)^{\frac{1}{m-1}}}$$

where n_j is a number of points in cluster *j*, *c* is a number of clusters.

2. Evaluate membership function using the distance between the object and the "trace" left by moving centroids can be used:

$$u_{ij} = \frac{1}{\sum_{k=1}^{c}\left(\dfrac{\|x_i - t_{c_j}\|}{\|x_i - t_{c_k}\|}\right)^{\frac{1}{m-1}}}$$

where t_{cj} is a nearest point of the "trace" left by moving centroid *j*.

Results

Testing of the algorithm was performed on Fisher's "Iris Data Set" [7].

The data set "Iris Data Set" contains 150 irises of three species, 50 of each. The object features here are the geometric characteristics. Four features were measured from each sample: the length and the width of the sepals and petals, in centimeters.

By using developed estimation procedure, it was found that there are 3 clusters in the data set. When clustering the "Iris Data Set" using developed algorithm the following result was obtained (Table 1).

Table 1: Cross table for clustering result given by developed algorithm

		Actual			Total
		0	1	2	
Result	0	36	0	0	36
	1	14	50	0	64
	2	0	0	50	50
Total		50	50	50	150
True positive, %		72	100	100	

As can be seen, the result of clustering was similar to the actual existing groups (Fig. 1).

The value of F_1-score was obtained using macro-averaging [8] and was 0.92. Below are the results obtained using classical *k*-means method (Table 2) and Ward's method hierarchical clustering (Table 3).

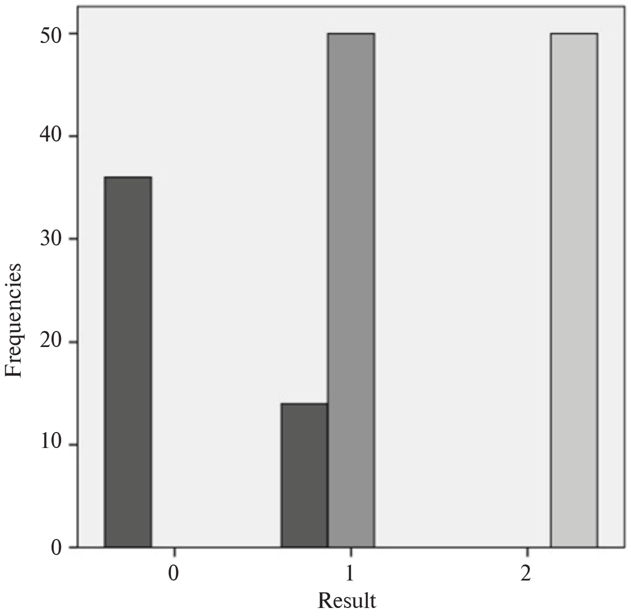

Figure 1: Diagram comparing clusters obtained using developed algorithm with actual classes: ■ — 0; ■ — 1; ■ — 2

Table 2: Cross-table for *k*-means clustering result

		Actual			Total
		0	1	2	
Result	0	42	0	10	52
	1	0	50	0	50
	2	8	0	40	48
Total		50	50	50	150
True positive, %		84	100	80	

Table 3: Cross table for clustering result obtained using Ward's method

		Actual			Total
		0	1	2	
Result	0	33	0	0	33
	1	0	50	0	50
	2	17	0	50	67
Total		50	50	50	150
True positive, %		66	100	100	

The result obtained with the *k*-means method has a F_1 value of 0.88. In general, the classic *k*-means showed the worst result on the test sample and gave more mixed clusters (Fig. 2).

The result obtained with Ward's method has a F_1 value of 0.90. This method formed clusters similar to those formed by developed method (Fig. 3).

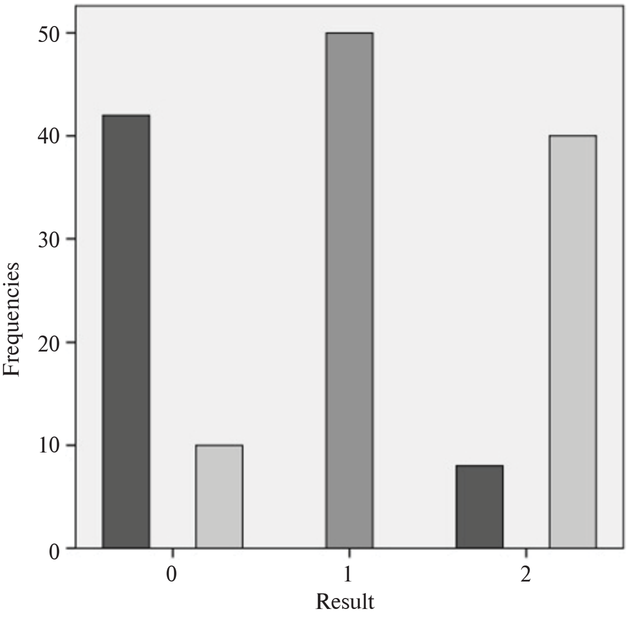

Figure 2: Diagram comparing clusters obtained using *k*-means with actual classes: ■ — 0; ■ — 1; ■ — 2

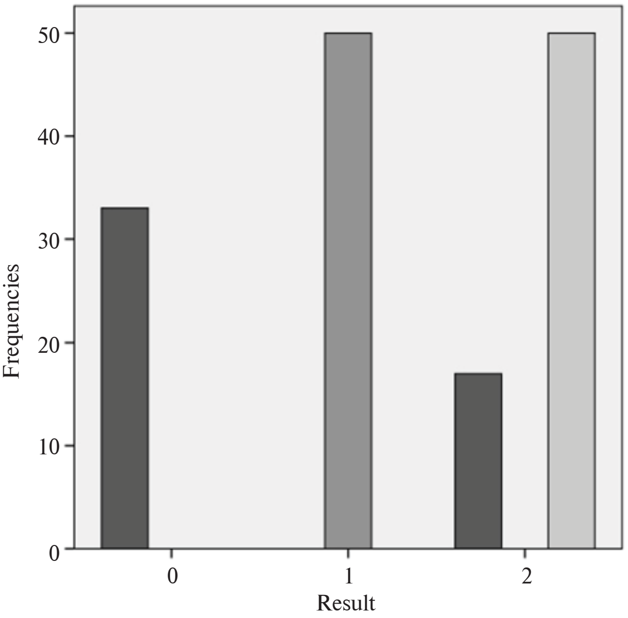

Figure 3: Diagram comparing clusters obtained using Ward's method with actual classes: ■ — 0; ■ — 1; ■ — 2

Discussion

As can be seen above, Ward's method gave similar results (see Fig. 3) to those obtained with *k*-means with a limited mass of working area (see Fig. 1). Some resemblance between the results of clustering is a consequence of the generation of hierarchical algorithms of non-spherical clusters, in the general case. The value of F_1-score for this result was 0.90, nevertheless, it is worse than for the development algorithm. This algorithm also tends to unite two closely situated classes in one big cluster.

In testing the algorithm, the calculation of the membership function using the minimum distance to the trace of the centroid of the working area of the algorithm was used, which provided the best result of clustering compared with the use of the mean distance from the studied point to all other cluster points.

Conclusions

The testing of the algorithms discussed in the article allows us to give preference to the developed version of fuzzy *k*-means with a limited mass of working area of cluster formation for cluster analysis problems with clusters of arbitrary form. The calculation of the membership function allows to obtain additional information about the structure of cluster entities, as well as to correct the result of clustering *k*-means with a limited mass, which is especially important for algorithms that obtain the result of clustering in one pass. Regarding the proximity of the qualitative results of the developed algorithm and Ward's method, it should be mentioned that the developed algorithm has a lower computational value since it does not require additional memory to store the matrix of distances and time for its recalculation. In addition, since the algorithm developed uses an a priori estimate of the clusters quantity, it has no problems related to the dendrogram cutting to get a result.

The developed algorithm will be further used in the development of cardiovascular state diagnostics system.

References

[1] Xiao Y, Yu J. Partitive clustering (K-means family). Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery. 2012;2(3):209-25. DOI: 10.1002/widm.1049

[2] Nastenko Y. The use of cluster analysis for partitioning mixtures of multidimensional functional characteristics of complex biomedical systems. J Automat Inform Sci. 1996;28(5-6):77-83. DOI: 10.1615/jautomatinfscien.v28.i5-6.100

[3] Boldak AA, Suharev DL. Determining number of clusters in statistical data. Visnyk NTUU KPI. Informatika, Upravlinnia ta Obchislyuvalna Tehnika. 2011;54(2):118-22.

[4] Shannon C. A mathematical theory of communication. Bell System Tech J. 1948;27(4):379-423, 623-656. DOI: 10.1002/j.1538-7305.1948.tb00917.x

[5] Kullback S, Leibler R. On information and sufficiency. Annals Math Stat. 1951;22(1):79-86. DOI: 10.1214/aoms/1177729694

[6] Scott D. On optimal and data-based histograms. Biometrika. 1979;66(3):605.

[7] Bezdek JC. Pattern recognition with fuzzy objective function algoritms. New York: Plenum Press; 1981. DOI: 10.1007/978-1-4757-0450-1

[8] Fisher R. UCI Machine learning repository: Iris data set [Internet]. Archive.ics.uci.edu. [cited 17 February 2018]. Available from: http://archive.ics.uci.edu/ml/datasets/Iris

[9] Asch V. Macro- and micro-averaged evaluation measures [Internet]. Clips.uantwerpen.be. 2012 [cited 13 April 2018]. Available from: https://www.clips.uantwerpen.be/~vincent/pdf/microaverage.pdf

В.С. Уманець, Б.О. Войник, В.А. Павлов, Є.А. Настенко

ОЦІНКА ЕФЕКТИВНОСТІ АЛГОРИТМІВ У ЗАДАЧІ КЛАСТЕРИЗАЦІЇ БІОЛОГІЧНИХ ОБ'ЄКТІВ

Проблематика. Завдання визначення функціонального зв'язку між біофізичними параметрами є складовою частиною актуальної проблеми пошуку оптимального впливу на біологічний об'єкт і на сьогодні не є повністю вирішеним. Однією з важливих задач у цій області є розбиття початкового простору ознак на такі області (кластери), які відносяться до різних функціональних співвідношень, що зв'язують біофізичні параметри, і які мають, у загальному випадку, довільну форму. Такі кластери в подальшому логічно називати функціональними. Для отримання й аналізу функціональних кластерів існує низка алгоритмів, кожен із яких має свої переваги й недоліки. У той же час розв'язання певної практичної задачі вимагає оцінки ефективності алгоритмів з точки зору адекватності виділення кластерів.

Мета. У статті для досить загального прикладу завдання кластеризації біологічних об'єктів (іриси Фішера) оцінюється ефективність низки типових інструментів кластеризації. Розглянуто застосування алгоритму *k*-середніх, алгоритму Варда, а також розробленої в роботі нечіткої версії кластеризації для алгоритму *k*-середніх з обмеженою масою робочої області формування кластерів.

Методика реалізації. В алгоритм включено процедуру апріорної оцінки кількості кластерів. Оцінка проводиться по гістограмі частот, для визначення оптимальної кількості стовпців гістограми обґрунтовується застосування формули Скотта. Алгоритм дає змогу формувати кластери довільної конфігурації з отриманням значення міри приналежності об'єкта кожному з кластерів. На наборі даних "Іриси Фішера" проведено порівняльне тестування зазначених алгоритмів.

Результати. Оптимальне значення F_1-score отримано для алгоритму, що запропонований у роботі – F_1 = 0,92, значення F_1 = 0,90 одержано для методу Варда і значення F_1 = 0,88 – для класичного алгоритму *k*-середніх.

Висновки. Отримані результати тестування свідчать, що в завданнях аналізу кластерів довільної форми доцільно віддати перевагу розробленій у дійсній роботі версії нечітких *k*-середніх з обмеженою масою робочої області формування кластерів. Розрахунок значення міри приналежності дає можливість в алгоритмі отримати додаткову інформацію про структуру кластерних утворень, а також здійснити поправки результату кластеризації *k*-середніх з обмеженою масою, що особливо важливо при формуванні кластерів за один прохід. Порівняння необхідних для розрахунку обчислювальних ресурсів для алгоритмів з відносно близькими результатами тесту також свідчить про перевагу запропонованого в роботі алгоритму. Порівняно з алгоритмом Варда йому необхідно менше обчислювальних ресурсів, оскільки не потрібна додаткова пам'ять для зберігання матриці відстаней і немає витрат часу на її перерахунок.

Ключові слова: кластеризація; *k*-середні; біологічний об'єкт; міра належності; оцінка кількості кластерів; нечітка кластеризація.

В.С. Уманец, Б.А. Войник, В.А. Павлов, Е.А. Настенко

ОЦЕНКА ЭФФЕКТИВНОСТИ АЛГОРИТМОВ В ЗАДАЧЕ КЛАСТЕРИЗАЦИИ БИОЛОГИЧЕСКИХ ОБЪЕКТОВ

Проблематика. Задача определения функциональной связи между биофизическими параметрами является составной частью актуальной проблемы поиска оптимального воздействия на биологический объект и в настоящее время не является полностью решенной. Одной из важных задач в этой области является разбиение исходного пространства признаков на такие области (кластеры), которые относятся к различным функциональным соотношениям, связывающим биофизические параметры, и имеют, в общем случае, произвольную форму. Такие кластеры в дальнейшем логично называть функциональными. Для получения и анализа функциональных кластеров существует ряд алгоритмов, каждый из которых обладает своими преимуществами и недостатками. В то же время решение определенной практической задачи требует оценки эффективности алгоритмов с точки зрения адекватности выделения кластеров.

Цель. В статье для достаточно общего примера задачи кластеризации биологических объектов (ирисы Фишера) оценивается эффективность ряда типичных инструментов кластеризации. Рассмотрено применение алгоритма *k*-средних, алгоритма Варда, а также разработанной в данной работе нечеткой версии кластеризации для алгоритма *k*-средних с ограниченной массой рабочей области формирования кластеров.

Методика реализации. В алгоритм включена процедура априорной оценки количества кластеров. Оценка проводится по гистограмме частот, для определения оптимального количества столбцов гистограммы обосновывается применение формулы Скотта. Алгоритм позволяет формировать кластеры произвольной конфигурации с получением значения меры принадлежности объекта каждому из кластеров. На наборе данных "Ирисы Фишера" проведено сравнительное тестирование указанных алгоритмов.

Результаты. Наилучшее значение F_1-score получено для алгоритма, предложенного в работе – F_1 = 0,92, F_1 = 0,90 для метода Варда и F_1 = 0,88 для классического алгоритма *k*-средних.

Выводы. Полученные результаты тестирования свидетельствуют о том, что в задачах анализа кластеров произвольной формы целесообразно отдать предпочтение разработанной в данной работе версии нечетких *k*-средних с ограниченной массой рабочей области формирования кластеров. Расчет значения меры принадлежности в алгоритме позволяет получить дополнительную информацию о структуре кластерных образований, а также осуществить поправки результата кластеризации *k*-средних с ограниченной массой, что особенно важно при формировании кластеров за один проход. Сравнение требуемых для расчета вычислительных ресурсов для алгоритмов с относительно близкими результатами теста также свидетельствует о преимуществе предложенного в работе алгоритма. По сравнению с алгоритмом Варда ему требуется меньше вычислительных ресурсов, так как не нужна дополнительная память для хранения матрицы расстояний и нет затрат времени на ее перерасчет.

Ключевые слова: кластеризация; *k*-средние; биологический объект; мера принадлежности; оценка количества кластеров; нечеткая кластеризация.