European Proceedings of
Computers and Technology
EpCT

www.europeanproceedings.com
e-ISSN: 2672-8834

DOI: 10.15405/epct.23021.27

HMMOCS 2022
International Workshop "Hybrid methods of modeling and optimization in complex systems"

APPROACH TO DATA CLUSTERING BASED ON MOLECULAR CHEMICAL REACTIONS WITH VARIOUS DISTANCE MEASURES

E. M. Markushin (a), G. Sh. Shkaberina (b), N. L. Rezova (c)*, L. A. Kazakovtsev (d)
*Corresponding author

(a) Reshetnev Siberian State University of Science and Technology, 31, Krasnoyarsky Rabochy Av., Krasnoyarsk, Russian Federation
(b) Laboratory “Hybrid Methods of Modeling and Optimization in Complex Systems”, Siberian Federal University, 79, Svobodny av., Krasnoyarsk, Russian Federation, Reshetnev Siberian State University of Science and Technology, 31, Krasnoyarsky Rabochy Av., Krasnoyarsk, Russian Federation, z_guzel@mail.ru
(c) Reshetnev Siberian State University of Science and Technology, 31, Krasnoyarsky Rabochy Av., Krasnoyarsk, Russian Federation, natalyakl@yandex.ru
(d) Reshetnev Siberian State University of Science and Technology, 31, Krasnoyarsky Rabochy Av., Krasnoyarsk, Russian Federation, Laboratory “Hybrid Methods of Modeling and Optimization in Complex Systems”, Siberian Federal University, 79, Svobodny av., Krasnoyarsk, Russian Federation, levk@bk.ru

Abstract

Automatic clustering involves dividing a set of objects into subsets so that the objects from one subset are more similar to each other than to the objects from other subsets according to some criterion. The paper proposes an algorithm for clustering data using the k-means algorithm combined with molecular chemical reactions and with various types of distance measures: Euclidean distance, Squared Euclidean distance, Manhattan distance. This approach mimics a chemical reaction process in which reactants interact with one another. Every chemical reaction process generate a new molecular structure in the environment. By molecular structure, we mean a possible solution to data clustering, by optimizing the molecular chemical reactions we mean optimizing the results of data clustering (search for a global optimal solution). The solution obtained with k-means is used as an initial molecular structure solution to optimize chemical reactions by generating new solutions: single-molecule collision, single-molecule decomposition, intermolecular collision, and intermolecular synthesis. Computational experiments demonstrate the comparative efficiency and accuracy of using the k-means algorithm combined with molecular chemical reactions.

Keywords: Data clustering, chemical reaction optimization, k-means

https://doi.org/10.15405/epct.23021.27
Corresponding Author: N. L. Rezova
Selection and peer-review under responsibility of the Organizing Committee of the conference
eISSN: 2672-8834

1. Introduction

One of the most well-known automatic grouping models is the k-means model (1) (Hossain et al., 2019; Li & Wu, 2012), which was proposed by Steinhaus (1956).

$$\text{argmin}F(X_1,\ldots,X_k) = \sum \underline{\min}_{j\in\{1,k\}} \left\|X_j - A_i\right\|^2 \qquad (1)$$

The goal of the k-means problem is to find k points (centers, centroids) $X_1,\ldots,X_k$ in a d-dimensional space such that the sum of the squared distances from the known points $A_1,\ldots,A_N$ to the nearest of the desired points reaches a minimum. Lloyd (1982) algorithmically implemented the k-means problem. In the works MacQueen (1967), the basic k-means algorithm, also known as Lloyd's algorithm, consists of iterative repetition of two steps:

- Given: k initial cluster centers (centroids).

Step 1. Create a new cluster Cj ($j = \overline{1,k}$) by assigning each data point to the closest cluster center (centroid).

Step 2. Calculation of new cluster centers.

Repeat steps 1 and 2 until there are no more changes within each cluster.

In the k-means algorithm, it is necessary to initially predict the number of groups (subsets). In addition, the result obtained depends on the initial choice of centers.

Kaufman and Rousseeuw (1987) presented the k-medoids model (PAM, Partitioning Around Medoids) close to k-means. The centers are clustered objects (medoids) that are part of the set under study. The algorithm is more resistant to outliers and noise than the k-means algorithm; however, it is inefficient when applied to large datasets due to time complexity. The k-medians algorithm (Jain & Dubes, 1988) is a variation of the k-means algorithm, where the median is calculated instead of the mean to determine the cluster centroid. Kaufman and Rousseeuw (1990) proposed the CLARA (Clustering Large Applications) algorithm, based on the PAM algorithm, for clustering objects in large databases in order to reduce computation time. In (Ng & Han, 2002) the authors propose the CLARANS (Clustering Large Applications based upon Randomized Search) algorithm, which is aimed at using a randomized search to facilitate the clustering of a large number of objects.

Algorithms k-means, k-medians, k-medoid, PAM, CLARA, CLARANS belong to the class of clustering algorithms based on partitioning methods (Partitioning clustering).

The current literature offers many heuristic approaches (Arthur & Vassilvitskii, 2007) to setting the initial centroids for the k-means algorithm, which are basically various evolutionary and random search methods. Local search algorithms and randomized algorithms based on them are presented in a large number of publications. For example, Variable Neighborhood Search (VNS) algorithms (Mladenović & Hansen, 1997; Rozhnov et al., 2019) or agglomeration algorithms (Sun et al., 2014) sometimes perform well. Initialization procedures for local search algorithms are also widely presented, including random filling and estimation of the distribution of demand points (Arthur & Vassilvitskii, 2007). However, in many cases, even repeated runs of simple local search algorithms from various randomly generated

https://doi.org/10.15405/epct.23021.27
Corresponding Author: N. L. Rezova
Selection and peer-review under responsibility of the Organizing Committee of the conference
eISSN: 2672-8834

solutions do not provide a solution to the problem close to the global optimum. More complex algorithms make it possible to obtain the values of the objective function (1) many times better than local search methods (Rozhnov et al., 2019).

A popular idea is to use genetic algorithms and other evolutionary approaches to improve local search results (Krishna & Murty, 1999; Maulik & Bandyopadhyay, 2000). Such algorithms combine local minima obtained using the k-means algorithm. Heretic algorithms operate on a certain set (population) of candidate solutions and include special genetic operators (algorithms) for initialization of selection, crossover, and mutation. The mutation operator randomly changes the resulting solutions and provides some diversity in the population. However, in such algorithms, as the number of iterations increases, the population degenerates into a certain set of solutions that are close to each other. Larger populations as well as dynamically growing populations improve this situation.

Classical clustering algorithms, such as k-means, perform a local search, improving the previous result. The result of the algorithm depends on the initial solution chosen. Therefore, the search for the optimal solution requires multiple attempts to run algorithms with procedures for random selection of initial solutions or random selection of local search (Vidyasagar, 1998), or the use of these procedures in the algorithm simultaneously. More complex algorithms can be applied, such as genetic algorithms, neural networks (Holland, 1975), simulated annealing algorithm (Kirkpatrick et al., 1983). These algorithms are based on the idea of modeling natural processes. The authors of the paper "Genetic K-Means algorithm" (Krishna & Murty, 1999) showed the advantage of genetic algorithms over classical algorithms. The genetic algorithm is a representative of evolutionary optimization algorithms (search for the optimal solution).

In his work, J. Holland introduces the concept of a fitness function (Holland, 1975), which is intended to determine the best solution for a genetic algorithm. The fitness function is the objective function. The purpose of crossing genes is to achieve the best value of the objective function.

The genetic algorithm for solving the discrete p-median problem proposed by Hosage and Goodchild (1986) preceded the genetic algorithm for the k-means problem. The authors of (Bozkaya et al., 2002) presented a genetic algorithm with solution coding in the form of a set of indices of network nodes chosen as a center using several crossing operators. The algorithm gave more accurate results with very slow convergence. In their work, Alp et al. (2003) presented a simple and faster genetic algorithm with a special crossing procedure - a greedy (agglomerative) heuristic procedure, which also gives exact results for the p-median network problem. This idea was inherited in (Kazakovtsev & Antamoshkin, 2014). Many mutation methods presented in (Kazakovtsev & Antamoshkin, 2014; Kwedlo & Iwanowicz, 2010) can be used in genetic algorithms for k-means and similar problems. In the k-means algorithm, usually the initial solution is a subset of the original data. In (He & Yu, 2019), the authors solve an alternative k-means problem aimed at increasing the stability of clustering instead of minimizing (1). The authors of (Pizzuti & Procopio, 2017) solve the problem with a mathematical formulation different from (1) and use cluster recalculation in accordance with (1) as a mutation operator. A similar coding is used in (Krishna & Murty, 1999), where the authors propose a mutation operator that changes the assignment of individual data objects to clusters. In the mutation operator is described as a procedure that guarantees the diversity of a population (Eremeev, 2012). For the k-means and p-median problem, the mutation procedure, as a rule, changes one or more solutions, replacing some centers (Krishna & Murty, 1999; Maulik & Bandyopadhyay, 2000).

Different mutation operators have been developed for different decision encoding methods: bit inversion for binary encoding (Holland, 1975), swap, insertion, inversion, and bias permutation (Larranaga et al., 1999) for variable length solutions, Gaussian mutation (Sarangi et al., 2015) and polynomial mutation for real encoding (Deb & Deb, 2012). Some studies propose a combination of mutation operators (Deep & Mebrahtu, 2011) or self-adaptive mutation operators (McGinley et al., 2011). The efficiency of various mutation operators depends on the parameters of the genetic algorithm (Osaba et al., 2014) and the type of problem (Karthikeyan et al., 2013). In Krishna and Murty (1999), Cheng et al. (2006), the authors propose to use the k-means algorithm as a mutation operator. Each of these algorithms declares a local search as a mutation operator. The structure of the genetic algorithm allows us to use a wide range of variants of genetic operators. However, local search is intended to improve an arbitrary solution by transforming it into a local optimum and thereby reducing rather than increasing the diversity of solutions.

2. Problem Statement

The problem of automatic classification (the problem of automatic clustering) can be described as follows. *N* objects of a certain set must be divided into *k* non-overlapping subsets so that the objects of one subset have similar features with each other and do not have them with objects from other subsets. The result clustering depends on the initially selected number of subsets and the measure of similarity (difference), expressed as a function of distances.

In Jain and Dubes (1988), the following formulation of the cluster analysis problem is proposed: let there be a sample of research objects $A=\{A_1,...,A_N\}$, where *N* is the sample size. It is required to form $k\geq2$ classes (groups of objects). The number of classes can be preselected or determined automatically. Each object is described using a set of *M* variables $Z_1,...,Z_M$. The set $Z=\{ Z_1,...,Z_M\}$ can include variables of different types.

3. Research Questions

The search for an object clustering algorithm that has both high accuracy and stability of the result, and at the same time high speed, is one of the problems of automatic grouping of objects. The presented work is devoted to the research and development of a new algorithm for automatic grouping of objects, which will improve the accuracy and stability of the result of solving practical problems.

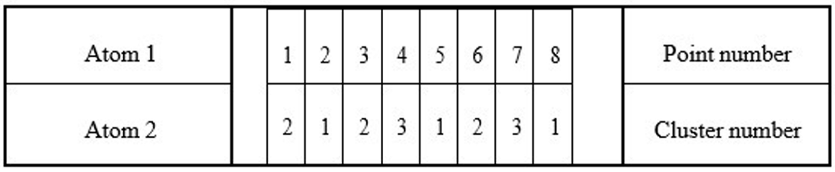
4. Purpose of the Study

The purpose of the study of this work is to improve the accuracy and stability of the result of solving problems of automatic grouping of objects.

5. Research Methods

In this paper, we propose an algorithm similar to the genetic algorithm, in which molecular chemical reactions are used as the mutation procedure (KCR algorithm). This approach mimics a chemical reaction process in which reactants interact with one another. Every chemical reaction process generate a new molecular structure in the environment (Pan et al., 2015). By molecular structure, we mean a possible solution to data clustering, by optimizing the molecular chemical reactions we mean optimizing the results of data clustering (search for a global optimal solution).

Imagine a possible solution to clustering as a molecular structure. Let each molecule consist of two sets of atoms. The first set of atoms contains a sequence of point numbers A = {A1, A2, … , AN}, the second set contains the numbers of clusters from X={X1, X2, ... , Xk} to which the corresponding points belong. Figure 1 shows an example of a possible molecular structure that contains three clusters and eight points.



Atom 1	1	2	3	4	5	6	7	8	Point number
Atom 2	2	1	2	3	1	2	3	1	Cluster number

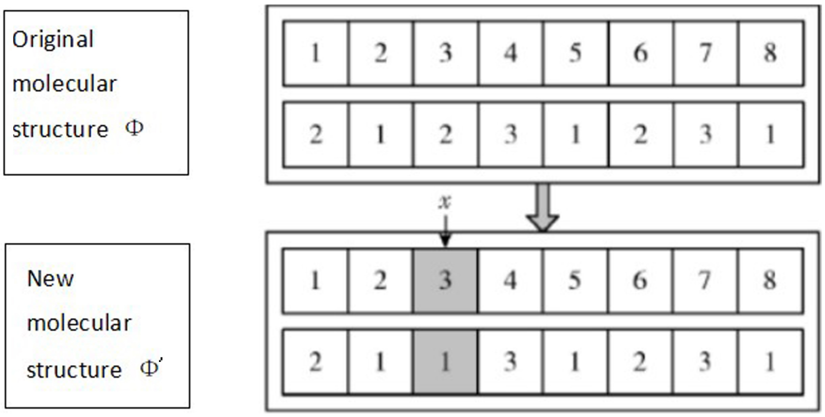
Figure 1. Chemical molecular structure coding

5.1. Generation of new clustering solutions

The paper proposes to consider and apply four types of chemical reactions: single- molecule collision, single-molecule decomposition, intermolecular collision and intermolecular synthesis (McNaught & Wilkinson, 1997).

5.1.1. Single molecule collision

In one substitution reaction, one element replaces another in a compound. The new molecular structure Φ' is obtained as follows. In the initial molecular structure, we randomly select a point and change its membership in the *X*={*X*₁, *X*₂, ... , *X*ₖ} cluster randomly (Figure 2).



Original molecular structure Φ

1	2	3	4	5	6	7	8
2	1	2	3	1	2	3	1

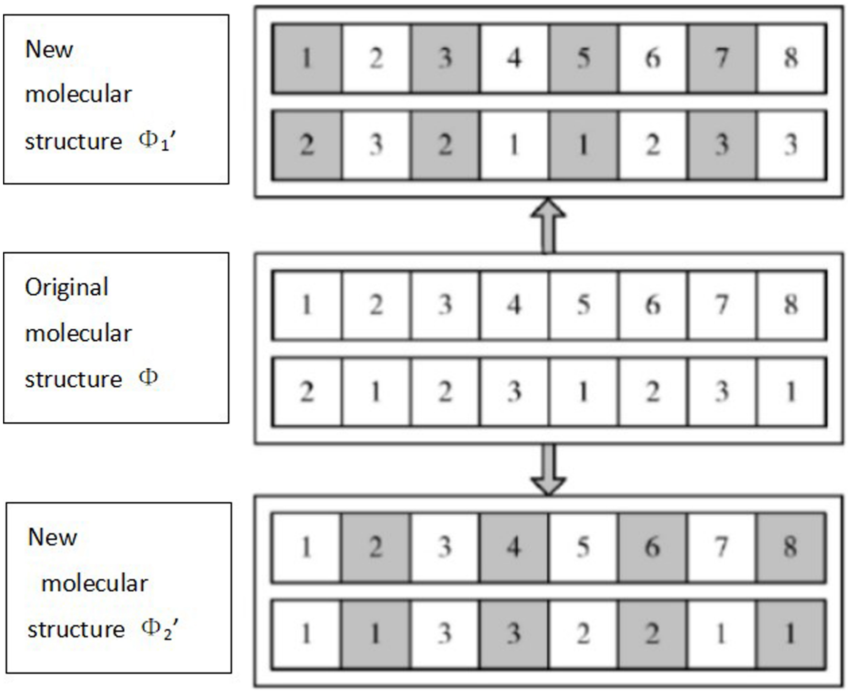
New molecular structure Φ′

1	2	3	4	5	6	7	8
2	1	1	3	1	2	3	1

Figure 2. Single molecule collision

5.1.2. Single molecule decomposition

A decomposition reaction is an approach where a more complex substance breaks down into simpler parts. In the clustering problem, two new molecular structures Φ1' and Φ2' are generated from the initial molecular structure (Figure 3). The new molecular structure Φ1' is obtained as follows. Points with uneven numbers and their corresponding cluster numbers are stored in the new molecular structure, and for points with even numbers, we determine their belonging to a cluster of $X=\{X_1, X_2, \dots , X_k\}$ randomly. The new molecular structure Φ2' is obtained as follows. Points with even numbers and the corresponding cluster numbers are stored in the new molecular structure, and for points with uneven numbers, we determine their belonging to a cluster of $X=\{X_1, X_2, \dots , X_k\}$ randomly.



New molecular structure Φ₁'

1	2	3	4	5	6	7	8
2	3	2	1	1	2	3	3

Original molecular structure Φ

1	2	3	4	5	6	7	8
2	1	2	3	1	2	3	1

New molecular structure Φ₂'

1	2	3	4	5	6	7	8
1	1	3	3	2	2	1	1

Figure 3. Single molecule decomposition

5.1.3. Intermolecular collision

In intermolecular collisions, the anions and cations of two compounds switch places and form two entirely different compounds. New molecular structures Φ1' и Φ2' is obtained as follows. The numbers of two points x and y ($x < y$) are randomly chosen in two initial molecular structures Φ1 and Φ2. The new molecular structure Φ1' is obtained by copying the initial molecular structure Φ2. Then we replace the cluster numbers for the ith points ($i \in [x,y]$) from the molecular structure Φ1. The new molecular structure Φ2' is obtained by copying the initial molecular structure Φ1. Then we replace the cluster numbers for the ith points ($i \in [x,y]$) from the molecular structure Φ2 (Figure 4).

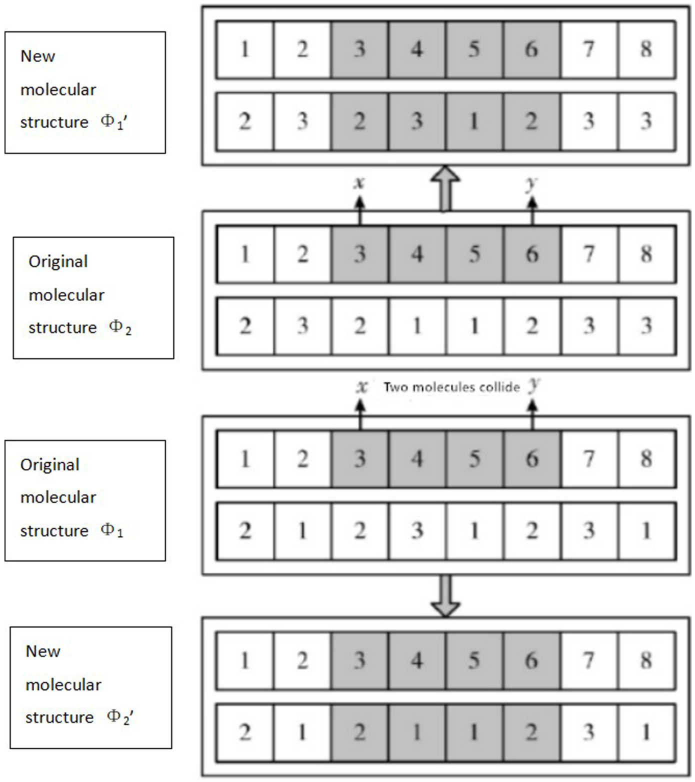

New molecular structure Φ₁′
1 2 3 4 5 6 7 8
2 3 2 3 1 2 3 3
x
y
Original molecular structure Φ₂
1 2 3 4 5 6 7 8
2 3 2 1 1 2 3 3
x Two molecules collide y
Original molecular structure Φ₁
1 2 3 4 5 6 7 8
2 1 2 3 1 2 3 1
New molecular structure Φ₂′
1 2 3 4 5 6 7 8
2 1 2 1 1 2 3 1

Figure 4. Intermolecular collision

5.1.4. Intermolecular synthesis

In a synthesis reaction, two or more simple substances combine to form a more complex substance. The number of point *x* is randomly chosen in two initial molecular structures Φ1 and Φ2. The new molecular structure Φ' is obtained by copying the initial molecular structure Φ1. Then we replace the cluster numbers for the *i*th points (i ∈ [*x*+1,*n*]) from the molecular structure Φ2 (Figure 5).

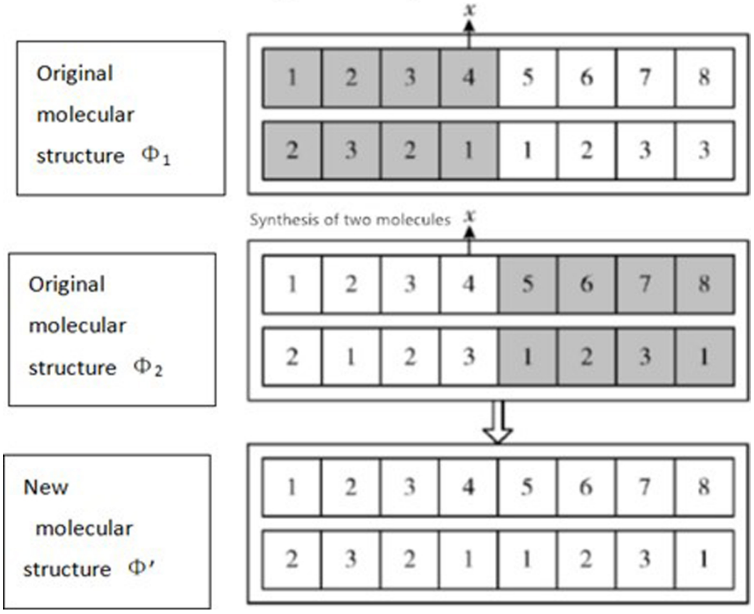

x
Original molecular structure Φ₁
1 2 3 4 5 6 7 8
2 3 2 1 1 2 3 3
Synthesis of two molecules x
Original molecular structure Φ₂
1 2 3 4 5 6 7 8
2 1 2 3 1 2 3 1
New molecular structure Φ′
1 2 3 4 5 6 7 8
2 3 2 1 1 2 3 1

Figure 5. Intermolecular synthesis

For each new solution (new molecular structure), we calculate the value of the objective function. If the objective function value of the new solution is better than the objective function value of the initial solution, then the new molecular structure is a valid solution (see Algorithm 1). In our work, the value of the objective function is defined as (1). The distance measures used are described below.

- The Minkowski function (Kulin & Kuenne, 1962):

$$d(x,y)=\left(\sum_{i=1}^{M}|x_i-y_i|^p\right)^{\frac{1}{p}}, \tag{2}$$

where *x* and *y* are input vectors of dimension *M*. For parameter *p*, the following statement ate true (proof): for $p \geq 1$ and $p = \infty$ the distance is a metric; for $p < 1$ the distance is not a metric.

- The Euclidean distance (EuD). For $p = 2$, the function (2) the takes the form of Euclidean distance:

$$d(x,y)=\sqrt{\sum_{i=1}^{M}(x_i-y_i)^2}. \tag{3}$$

- The squared Euclidean distance (SEuD):

$$d(x,y)=\sum_{i=1}^{M}(x_i-y_i)^2. \tag{4}$$

- The Manhattan distance (ManD). For $p = 1$, we get (2) the Manhattan distance, which is the second most popular distance:

$$d(x,y)=\sum_{i=1}^{M}|x_i-y_i|. \tag{5}$$

5.2. KCR algorithm

Algorithm 1 KCR

Require: The initial data points A1,…,AN, numbers of clusters k, the number of iterations of chemical reaction process n.

Step 1. Generation of a random initial solution Φ = {X1 ... Xk};

Step 2. Apply the k-means algorithm to Φ to obtain a local optimum Φ′;

Step 3. i=1;

Step 4. Apply the procedures of molecular chemical reactions for the individual Φ′ to obtain a new solution Φ″:

Step 4.1. The generation of new clustering solution by Single molecule collision method from Φ′. Calculate the objective function (1) of a new molecule $F(\Phi_1'')$;

Step 4.2. The generation of new clustering solution by Single molecule decomposition method. Calculate the objective function (1) of a new molecule $F(\Phi_2'')$;

Step 4.3. Randomly the generation of new clustering solution $\Psi = \{X_1 \dots X_k\}$. The generation of new clustering solution by Intermolecular collision method. Calculate the objective function (1) of a new molecule $F(\Phi_3'')$;

Step 4.4. Randomly the generation of new clustering solution $\Upsilon = \{X_1 \dots X_k\}$. The generation of new clustering solution by Intermolecular synthesis method. Calculate the objective function (1) of a new molecule $F(\Phi_4'')$;

Step 5. Determine Φ'', where $F(\Phi'') = \min\,(F(\Phi_1''), F(\Phi_2''), F(\Phi_3''), F(\Phi_4''))$;

Step 6. Apply the k-means algorithm to Φ'' to obtain a local optimum Φ''';

Step 7. i=i+1;

Step 8. IF F(Φ''')>F(Φ') AND i<=n THEN $\Phi' \leftarrow \Phi''$; go to ШАГ 4 ELSE $\Phi' \leftarrow \Phi'''$

Step 9. Decode clustering solution Φ'.

The firstly, the k-means clustering algorithm searches for a local optimal solution Φ'. Next, the chemical reactions algorithm runs. Then the best solution is determined from those obtained using the chemical reactions algorithm Φ''. And again applying k-means clustering algorithm searches for a new solution Φ'''. If the found solution Φ''' improves the solution Φ', then the desired solution is the solution Φ''', otherwise we apply the procedures of molecular chemical reactions for the mutated individual Φ''.

6. Findings

In this section, we compare the results of the experiment performed with k-means and KCR algorithms with various types of distance measures: Euclidean distance, Squared Euclidean distance, Manhattan distance.

For the experiments, we used Synthetic (artificial) datasets (with cluster labels) (GitHub Inc. [GHI], 2022), as well as from samples of industrial products (Kazakovtsev et al., 2015; Shkaberina et al., 2020).

a. Square3 is artificial dataset contains the collection of four clusters (1000 data points, 2 dimensions).

b. Microchips 1526IE10_002 set of results of test effects on electrical and radio products for monitoring the current-voltage characteristics of input and output circuits of microcircuits (3987 data points, 67 dimensions).

Algorithms were implemented in Python. For the computational experiments, we used the following test system: AMD Ryzen 5 3500U, 2.10 GHz, 8 CB RAM.

For various datasets, the minimum (min), maximum (max), mean (mean), standard deviation (σ), coefficient of variation (V) and the the span factor (R) of the objective function are calculated (Tables 1, 2). In the tables, the best mean values of objective function are given in bold.

Table 1. Objective function value summarized after 30 attempts. Synthetic (artificial) dataset

Parameter	k-means			KCR		
	EuD	SEuD	ManD	EuD	SEuD	ManD
min	100.2318	100.2318	102.5189	100.2318	100.2318	102.4449
max	100.2379	100.2379	160.385	100.2379	100.2369	159.8675
mean	100.2351	100.2353	115.9838	100.2334	100.2334	115.8819
σ	0.003029	0.002945	24.29265	0.001683	0.001787	24.15255
V	0.003021	0.002938	20.94486	0.001679	0.001783	20.84239
R	0.006071	0.006071	57.86609	0.006071	0.00513	57.42261

Table 2. Objective function value summarized after 30 attempts. Microchips 1526IE10_002

Parameter	k-means			KCR		
	EuD	SEuD	ManD	EuD	SEuD	ManD
Two-batch mixed lot (197 data points, 41 dimensions)						
min	0.026981824	0.026981824	0.027030361	0.026976786	0.026976786	0.026981824
max	0.026981824	0.026981824	0.027030361	0.026981824	0.026981824	0.027030361
mean	0.026981824	0.026981824	0.027030361	**0.026979305**	0.026980648	0.027027312
σ	0	0	0	2.52E-06	2.13E-06	9.20E-06
V	0	0	0	0.00933682	0.007897671	0.034030838
R	0	0	0	5.04E-06	5.04E-06	4.85E-05
Three-batch mixed lot (300 data points, 41 dimensions)						
min	0.048329262	0.048329262	0.048419577	0.048329262	0.048329262	0.048358367
max	0.048332067	0.048332067	0.04847852	0.048329262	0.048329262	0.048476331
mean	0.048329449	0.048329636	0.048465557	**0.048329262**	**0.048329262**	0.048460111
σ	7.00E-07	9.53E-07	1.98E-05	3.46E-18	3.46E-18	3.34E-05
V	0.001447527	0.001972639	0.040875312	7.16E-15	7.16E-15	0.068835581
R	2.80E-06	2.80E-06	5.89E-05	6.94E-18	6.94E-18	0.000117964
Four-batch mixed lot (446 data points, 62 dimensions)						
min	0.011871273	0.011871273	0.012025967	0.011871273	0.011871273	0.012021142
max	0.016308472	0.016308472	0.016554985	0.011885287	0.011904385	0.013707829
mean	0.012022907	0.012906738	0.012343073	**0.011873588**	0.011874292	0.012097325
σ	0.000795859	0.001876659	0.001125694	4.39E-06	8.41E-06	0.000299143
V	6.619519898	14.54015291	9.12004666	0.0369922	0.070829331	2.472800111
R	0.0044372	0.0044372	0.004529018	1.40E-05	3.31E-05	0.001686687
Full mixed lot (3987 data points, 67 dimensions)						
min	0.118946137	0.118946137	0.119723176	0.118946137	0.118946575	0.1197271
max	0.164975905	0.157709299	0.159640787	0.143888254	0.157709241	0.14263887
mean	0.129590909	0.129384046	0.126030278	0.125914328	0.127245642	**0.124172035**
σ	0.012357251	0.011716148	0.010097741	0.008068167	0.010446673	0.007021172
V	9.53558456	9.055326403	8.01215488	6.407664014	8.209847419	5.654390752
R	0.046029769	0.038763163	0.039917611	0.024942118	0.038762666	0.02291177

Computational experiments showed that in the vast majority of cases, minimal mean objective function value was demonstrated by KCR algorithm with Euclidean distance. However, using the KCR algorithm with Manhattan distance, in most cases, improves the accuracy of data clustering (Figure 6). In addition, the clustering accuracy increases with increasing number of points in the dataset (Table 3).

Table 3. Accuracy of data clustering with various measure of distance

Dataset	k-means			KCR		
	EuD	SEuD	ManD	EuD	SEuD	ManD
Synthetic dataset	0.771	0.772	0.792	0.772	0.772	0.793
Two-batch mixed lot	0.754	0.754	0.763	0.75	0.75	0.759
Three-batch mixed lot	0.695	0.695	0.701	0.698	0.698	0.704
Four-batch mixed lot	0.676	0.609	0.687	0.678	0.678	0.690
Full mixed lot	0.493	0.493	0.515	0.509	0.509	0.515

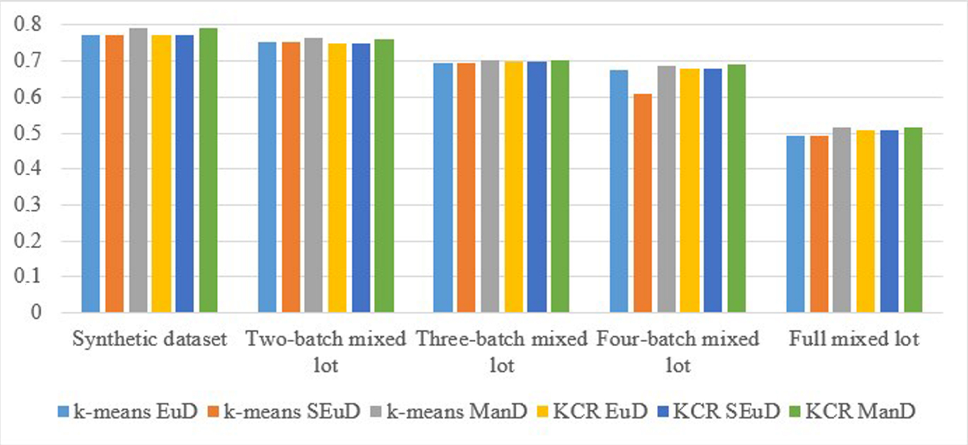


Figure 6. Accuracy of data clustering with various measure of distance

7. Conclusion

We proposed KCR algorithm for data clustering, which combines k-means algorithm and chemical reaction algorithms, could achieve an effective balance between local search capabilities and global exploration capabilities. The new algorithm improves the accuracy of solving the k-means problem.

Computational experiments showed that in the vast majority of cases, minimal mean objective function value was demonstrated by KCR algorithm with Euclidean distance. However, using the KCR algorithm with Manhattan distance, in most cases, improves the accuracy of data clustering. In addition, the clustering accuracy increases with increasing number of points in the dataset.

Acknowledgments

This work was supported by the Ministry of Science and Higher Education of the Russian Federation (Grant No.075-15-2022-1121).

References

Alp, O., Erkut, E., & Drezner, Z. (2003). An Efficient Genetic Algorithm for the p-Median Problem. *Annals of Operations Research*, *122*(1-4), 21-42. https://doi.org/10.1023/A:1026130003508

https://doi.org/10.15405/epct.23021.27
Corresponding Author: N. L. Rezova
Selection and peer-review under responsibility of the Organizing Committee of the conference
eISSN: 2672-8834

Arthur, D., & Vassilvitskii, S. (2007). K-Means++: The Advantages of Careful Seeding. *Proceedings of SODA'07, SIAM*, 1027-1035.

Bozkaya, B. A., Zhang, J., & Erkut, E. (2002). Genetic Algorithm for the p-Median Problem. *Facility Location: Applications and Theory* (pp. 179-205). Springer. https://doi.org/10.1007/978-3-642-56082-8_6

Cheng, S. S., Chao, Y. H., Wang, H. M., & Fu, H. C. (2006). A prototypes-embedded genetic k-means algorithm. In *18th International Conference on Pattern Recognition (ICPR'06)* (Vol. 2, pp. 724-727). IEEE. https://doi.org/10.1109/ICPR.2006.155

Deb, D., & Deb, K. (2012). Investigation of mutation schemes in real-parameter genetic algorithms. In *Swarm, Evolutionary, and Memetic Computing: Third International Conference, SEMCCO 2012, Bhubaneswar, India, December 20-22, 2012. Proceedings 3* (pp. 1-8). Springer. https://doi.org/10.1007/978-3-642-35380-2_1

Deep, K., & Mebrahtu, H. (2011). Combined mutation operators of genetic algorithm for the travelling salesman problem. *International Journal of Combinatorial Optimization Problems and Informatics, 2*(3), 1-23. https://www.redalyc.org/pdf/2652/265219635002.pdf

Eremeev, A. V. (2012). A genetic algorithm with tournament selection as a local search method. *Journal of Applied and Industrial Mathematics, 6*(3), 286-294. https://doi.org/10.1134/s1990478912030039

GitHub Inc. (2022, November 18). *Artificial Clustering Datasets.* https://github.com/milaan9/Clustering-Datasets (accessed on 14 October 2022).

He, Z., & Yu, C. (2019). Clustering stability-based Evolutionary K-Means. *Soft Computing, 23*(1), 305-321. https://doi.org/10.1007/s00500-018-3280-0

Holland, J. (1975). *Adaptation in Natural and Artificial Systems*. University of Michigan Press.

Hosage, C. M., & Goodchild, M. F. (1986). Discrete Space Location-Allocation Solutions from Genetic Algorithms. *Annals of Operations Research*, *6*, 35-46. https://doi.org/10.1007/bf02027381

Hossain, M. Z., Akhtar, M. N., Ahmad, R. B., & Rahman, M. (2019). A dynamic K-means clustering for data mining. *Indonesian Journal of Electrical Engineering and Computer Science, 13*(2), 521. https://doi.org/10.11591/ijeecs.v13.i2.pp521-526

Jain, A. K., & Dubes, R. C. (1988). *Algorithms for Clustering Data*. Prentice-Hall.

Karthikeyan, P., Baskar, S., & Alphones, A. (2013). Improved genetic algorithm using different genetic operator combinations (GOCs) for multicast routing in ad hoc networks. *Soft Computing, 17,* 1563-1572. https://doi.org/10.1007/s00500-012-0976-4

Kaufman, L., & Rousseeuw, P. J. (1987). Clustering by means of medoids. In *Proc. Statistical Data Analysis Based on the L1 Norm Conference* (pp. 405-416). Springer.

Kaufman, L., & Rousseeuw, P. J. (1990). *Finding groups in data: an introduction to cluster analysis*. Wiley.

Kazakovtsev, L. A., & Antamoshkin, A. N. (2014). Genetic Algorithm with Fast Greedy Heuristic for Clustering and Location Problems. *Informatica*, *38*(3), 229-240.

Kazakovtsev, L. A., Antamoshkin, A. N., & Masich, I. S. (2015). Fast deterministic algorithm for EEE components classification. *2015 IOP Conference Series: Materials Science and Engineering*, *94*, 012015. https://doi.org/10.1088/1757-899x/94/1/012015

Kirkpatrick, S., Gelatt, C. D., & Vecchi, M. P. (1983). Optimization by simulated annealing. *Science*, *220*(4598), 671-680. https://doi.org/10.1126/science.220.4598.671

Krishna, K., & Murty, M. M. (1999). Genetic K-Means algorithm. IEEE *Transactions on Systems. Part B (Cybernetics)*, *29*(3), 433-439. https://doi.org/10.1109/3477.764879

Kulin, H. W., & Kuenne, R. E. (1962). An efficient algorithm for the numerical solution of the generalized Weber problem in spatial economics. *Journal of Regional Science, 4*(2), 21-33.

Kwedlo, W., & Iwanowicz, P. (2010). Using Genetic Algorithm for Selection of Initial Cluster Centers for the K-Means Method. *ICAISC 2010: Artificial Intelligence and Soft Computing*, 165-172. https://doi.org/10.1007/978-3-642-13232-2_20

Larranaga, P., Kuijpers, C. M. H., Murga, R. H., Inza, I., & Dizdarevic, S. (1999). Genetic algorithms for the travelling salesman problem: A review of representations and operators. *Artificial intelligence review, 13*, 129-170. https://doi.org/10.1023/A:1006529012972

Li, Y., & Wu, H. (2012). A Clustering Method Based on K-Means Algorithm. *Physics Procedia, 25,* 1104-1109. https://doi.org/10.1016/j.phpro.2012.03.206

Lloyd, S. P. (1982). Least Squares Quantization in PCM. *IEEE Transactions on Information Theory*, *28*, 129-137. https://doi.org/10.1109/tit.1982.1056489

MacQueen, J. B. (1967). Some Methods of Classification and Analysis of Multivariate Observations. *Proceedings of the 5th Berkley Symposium on Mathematical Statistics and Probability*, *1*, 281-297.

Maulik, U., & Bandyopadhyay, S. (2000). Genetic Algorithm-Based Clustering Technique. *Pattern Recognition*, *33*(9), 1455-1465. https://doi.org/10.1016/s0031-3203(99)00137-5

McGinley, B., Maher, J., O'Riordan, C., & Morgan, F. (2011). Maintaining healthy population diversity using adaptive crossover, mutation, and selection. *IEEE Transactions on Evolutionary Computation, 15*(5), 692-714. https://doi.org/10.1109/tevc.2010.2046173

McNaught, A. D., & Wilkinson, A. (1997). *Compendium of Chemical Terminology* (2nd ed., the "Gold Book"). Blackwell Scientific Publications.

Mladenović, N., & Hansen, P. (1997). Variable neighborhood search. *Computers & operations research, 24*(11), 1097-1100. https://doi.org/10.1016/S0305-0548(97)00031-2

Ng, R. T., & Han, J. (2002). CLARANS: a method for clustering objects for spatial data mining. *IEEE Transactions on Knowledge and Data Engineering, 14*(5), 1003-1016. https://doi.org/10.1109/tkde.2002.1033770

Osaba, E., Carballedo, R., Diaz, F., Onieva, E., de la Iglesia, I., & Perallos, A. (2014). Crossover versus Mutation: A Comparative Analysis of the Evolutionary Strategy of Genetic Algorithms Applied to Combinatorial Optimization Problems. *The Scientific World Journal*, 1-22. https://doi.org/10.1155/2014/154676

Pan, G., Xu, Y., Ouyang, A., & Zheng, G. (2015). An Improved Artificial Chemical Reaction Optimization Algorithm for Job Scheduling Problem in Grid Computing Environments. *Journal of Computational and Theoretical Nanoscience, 12*(7), 1300-1310. https://doi.org/10.1166/jctn.2015.3890

Pizzuti, C., & Procopio, N. (2017). A K-means based genetic algorithm for data clustering. In *International Joint Conference SOCO'16-CISIS'16-ICEUTE'16: San Sebastián, Spain, October 19th-21st, 2016 Proceedings 11* (pp. 211-222). Springer, Cham. https://doi.org/10.1007/978-3-319-47364-2_21

Rozhnov, I. P., Orlov, V. I., & Kazakovtsev, L. A. (2019). VNS-Based Algorithms for the Centroid-Based Clustering Problem. *Facta Universitatis Series: Mathematics and Informatics*, *34*(5), 957-972. https://doi.org/10.22190/fumi1905957r

Sarangi, A., Lenka, R., & Sarangi, S. K. (2015). Design of linear phase FIR high pass filter using PSO with gaussian mutation. *Swarm, Evolutionalry, and Memetic Computing, SEMCCO 2014*, *8947*, 471-479. https://doi.org/10.1007/978-3-319-20294-5_41

Shkaberina, G. S., Orlov, V. I., Tovbis, E. M., & Kazakovtsev, L. A. (2020). On the Optimization Models for Automatic Grouping of Industrial Products by Homogeneous Production Batches. *Communications in Computer and Information Science*, *1275*, 421-436. https://doi.org/10.1007/978-3-030-58657-7_33

Steinhaus, H. (1956). Sur la divisiondes corps materiels en parties. *Bulletin L'Académie Polonaise des Science, 3*(4), 801-804.

Sun, Z., Fox, G., Gu, W., & Li, Z. (2014). A parallel clustering method combined information bottleneck theory and centroid-based clustering. *The Journal of Supercomputing, 69*, 452-467. https://doi.org/10.1007/s11227-014-1174-1

Vidyasagar, M. (1998). Statistical learning theory and randomized algorithms for control. *IEEE Control Systems, 12*, 69-85. https://doi.org/10.1109/37.736014