On the application of one approach for data clustering in the agro-industrial complex

Ivan P. Rozhnov[1,2*], *Lev* A. Kazakovtsev[1,2], *Margarita* V. Karaseva[1,2], *Natalya* L. Rezova[2], and *Igor* A. Gaiduk[2]

[1]Siberian Federal University, 79, Svobodny av., Krasnoyarsk, 660041, Russian Federation
[2]Reshetnev Siberian State University of Science and Technology, 31, Krasnoyarsky Rabochy av., Krasnoyarsk, 660037, Russian Federation

Abstract. The paper presents an approach to the automatic grouping algorithms development based on parametric optimization models for processing high-volume data in the agrarian and industrial complex. Combined search algorithms with alternating randomized neighborhoods show much more stable results (give a smaller minimum value, and also have a low standard deviation of the target function) and hence better performance compared to known (so-called classical) algorithms, such as j-means and k-means.

1 Introduction

Nowadays, the increasing use of big data stimulates the development and implementation of tools and methods of analysis and processing of large volumes of data in various fields and areas of life, including industry, energy, medicine, civil defense and emergency situations, logistics, public administration, etc. These industries are successfully developing along with the progress in the field of digital technologies. However, although the agrarian and industrial complex (AIC) of Russia is among the main world industries, it doesn't widely apply the newest methods of agricultural production and digital technologies [1, 2].

There is an increase in the volume of data. It is necessary to point out also an increase in the quantity and quality of modern tools and solutions at the present stage of the agrarian and industrial complex development. So, one can observe the need for reliable findings for making managerial decisions [1-3].

The process of digitalization in the agrarian and industrial complex is also an opportunity to create the complex production and logistics chains, covering agricultural producers and their suppliers, logistics, wholesale and retail companies into a single complex with predictive management. It will significantly reduce the cost of agricultural products, thus increasing the volume of production and sales, and the availability of food products for consumers [2].

* Corresponding author: ris2005@mail.ru

2 Digital technologies in the agrarian and industrial complex

It is possible to widely use such end-to-end digital technologies in the agrarian and industrial complex as:

- "artificial intelligence" - to create expert systems on soils, fertilizer systems, on the recognition of diseases and the development of recommendations on means of protection;
- "big data" - for analyzing field histories, physical and chemical composition of soils, land transactions and marketing research of agricultural machinery and equipment supplies, fertilizers, protection means, evaluating the impact of feeding rations on the productivity of farm animals of various breeds, as well as soil and weather conditions on crop variety yields [2].

Despite the importance of agriculture in the development of the state, it is characterized by low technical staffing and insufficient pace of infrastructure development in the field of information and scientific support [2]. By the order of the Government of the Russian Federation from 12.04.2020 N 993-r "Strategy of development of agro-industrial and fishery complex of the Russian Federation for the period till 2030" was approved. The Strategy stipulates that the introduction of digital technologies in the agro-industrial complex will help to improve the efficiency of work.

For example, potential nutrient deficiencies in the soil can be estimated applying artificial intelligence methods. The system will track changes in plants that are affected by soil defects and pests or plant spread across the field diseases. Having analyzed the problem, the system recommends methods of soil restoration to the agricultural producer, as well as other solutions that improve the quality and quantity of the crop. However, the concrete results of such studies in the agrarian and industrial complex are still negligible. Additional study of the technology application is required to understand the effectiveness and efficiency for real-life applications, including as a predictive control.

3 Approach to the clustering algorithms development

One of the promising fields in big data analytics (Big Data) is cluster analysis. Its range of application is very wide and it is used to solve problems in almost all areas of human life [4, 5].

The combined application of the method of greedy heuristics [4] with VNS-algorithms [6, 7] for problems of k-means [8, 9], k-medoid [10] and the CEM algorithm [3] was previously considered in detail in the works [3, 4]. The approach was applied to the development of clustering algorithms to improve the accuracy of results for automatic grouping algorithms. It is based on parametric optimization models with the combined application of search algorithms with alternating randomized neighborhoods and greedy agglomerative heuristic procedures [4]. Figure 1 presents a general scheme of this method.

4 Results of computational experiments

The authors used information from the open data of the Ministry of Agriculture of the Russian Federation for the study. They are "Catalogue of pesticides registered on the territory of the Russian Federation" (Table 1) and "Information on the state and use of lands of the Russian Federation by land" (Table 2).

Each of the algorithms was run 30 times when clustering data subsets. According to the results of our experiments for each algorithm we calculated the values of the target function: the minimum value (Minimum), the maximum value (Maximum), average value (Average)

and the standard deviation (SD). Algorithms k-means and j-means were run in multistart mode (Table 1 and 2). The best values are in bold.

Table 1. Results on the database "Catalogue of pesticides registered in the territory of the Russian Federation".

Algorithm	Value of the target function			
	Minimum	**Maximum**	**Average**	**SD**
k-means	3743.40	3744.62	3743.39	0.9346
j-means	3742.07	3743.52	3742.57	0.4487
GH-VNS1	**3741.97**	3743.08	3742.36	0.4020
GH-VNS2	**3741.97**	3743.15	**3742.06**	0.5028
GH-VNS3	**3741.97**	3742.10	**3741.99**	**0.0424**
GAGH+LS	3742.10	3745.73	3743.72	1.2199
GA FL	3741.99	3742.34	3742.10	0.2045
GA classical	3742.09	3742.88	3742.45	0.3489

Table 2. Results on the database "Information on the state and use of lands of the Russian Federation".

Algorithm	Value of the target function			
	Minimum	**Maximum**	**Average**	**SD**
k-means	53675.96	53681.52	53678.74	1.4126
j-means	53675.90	53684.88	53679.77	2.8062
GH-VNS1	**53671.89**	53671.89	**53671.89**	**0.0000**
GH-VNS2	**53672.24**	53674.44	**53673.34**	1.0476
GH-VNS3	**53672.84**	53675.76	**53674.30**	1.5916
GAGH+LS	53678.79	53693.63	53687.01	4.5961
GA FL	53708.14	53736.26	53716.26	8.4025
GA classical	53703.31	53724.42	53715.80	6.1660

The following abbreviations are used in tables as GH-VNS is clustering algorithm developed using the approach under consideration, GA is genetic algorithm, GAGH + LS is GA with a greedy heuristic procedure with local search and a real alphabet, GA FL is GA with recombination of subsets of the fixed length [5].

5 Conclusions

Taking into account the results of computational experiments, one can notice that the GH-VNS algorithms developed applying the presented approach (Figure 1) have more accurate (the value of the target function is less by the average) and more stable (lower standard deviation of the objective function) indicators compared to the considered classical algorithms (k- means, j-means, PAM, CEM), as well as some genetic algorithms. Thus, one can apply the approach considered in this paper to the processing of big data in the agrarian and industrial complex.

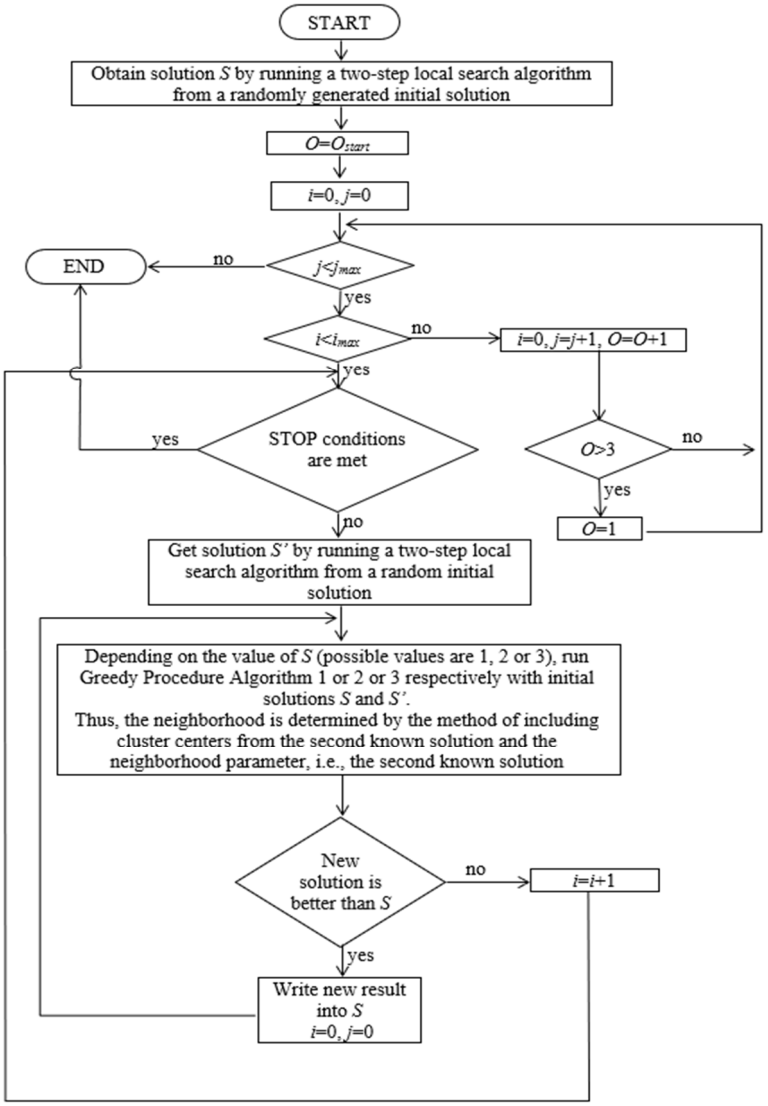

Fig. 1. General scheme of the method to the clustering algorithms development.

Simultaneously, it is desirable to apply several methods of evaluation to obtain more accurate results, taking into account the complexity of automatic grouping of objects to estimate the quality of clustering algorithms. On their basis one can check the result of the study of the certain clustering algorithms. At present, we observe the use of collective (ensemble) approaches, since automatic grouping algorithms are not universal [1]. This makes it possible to obtain a more accurate, stable solution and to reduce the dependence of the final solution on the initial given parameters of the algorithms.

The work was carried out within the framework of the state support program for leading scientific schools (grant of the President of the Russian Federation NSh-421.2022.4).

References

1. I. P. Rozhnov, L. A. Kazakovtsev, M. V. Karaseva, A. A. Stupina, E. V. Lapunova, IOP Conf. Series: Materials Science and Engineering **862**, (2020). https//doi.org/10.1088/1757-899X/862/4/042017
2. A. A. Shuvalov, Vestnik nauki **7(15)**, 91-96 (2019)
3. I. Rozhnov, L. Kazakovtsev, E. Bezhitskaya, S. Bezhitskiy, IOP Conf. Series: Materials Science and Engineering **537**, (2019). https//doi.org/10.1088/1757-899X/537/5/052032
4. L. Kazakovtsev, I. Rozhnov, Informatica **44**, 55-61 (2020). https//doi.org/10.31449/inf.v44i1.2737
5. Z. Drezner, H. Hamacher, *Facility location: applications and theory* (Berlin, Springer-Verlag, 2004), 460
6. Y. Kochetov, E. Alekseeva, T. Levanova, M. Loresh, Yugoslav Journal of Operations Research **15(1)**, 53-63 (2005)
7. P. Hansen, N. Mladenovic, D. Perez-Brito, J. Heuristics **4**, 335-350 (2001)
8. Z. Drezner, IMA Journal of Management Mathematics **26**, 1-9 (2013). https//doi.org/10.1093/imaman/dpt019
9. J. B. MacQueen, In Proceedings of the 5th Berkley Symposium on Mathematical Statistics and Probability, USA **1**, 281-297 (1965)
10. F. Garcia-Lopez, B. Melian-Batista, J. Moreno-Perez, M. Moreno-Vega, Journal of Heuristics **8**, 375-88 (2002). https://doi.org/10.1023/A:1015013919497