УДК 519.6

О ПРИМЕНЕНИИ ОДНОГО ПОДХОДА ДЛЯ ОБРАБОТКИ ДАННЫХ БОЛЬШОЙ РАЗМЕРНОСТИ

Рожнов Иван Павлович, кандидат технических наук,
старший научный сотрудник института информатики и телекоммуникаций
Сибирский государственный университет науки и технологий им. академика М.Ф. Решетнёва, Красноярск, Россия
e-mail: ris2005@mail.ru
Казаковцев Лев Александрович, доктор технических наук, доцент,
профессор кафедры «Информационные технологии и математическое обеспечение информационных систем», Институт экономики и управления АПК
Красноярский государственный аграрный университет, Красноярск, Россия
Сибирский государственный университет науки и технологий им. академика М.Ф. Решетнёва, Красноярск, Россия
e-mail: levk@bk.ru
Резова Наталья Леонидовна,
доцент кафедры «Информационно-управляющих систем», Институт информатики и телекоммуникаций
Сибирский государственный университет науки и технологий им. академика М.Ф. Решетнёва, Красноярск, Россия
e-mail: natalyakl@yandex.ru

Аннотация. Для обработки данных большого объема в агропромышленном комплексе показано применение подхода к разработке алгоритмов автоматической группировки, основанных на параметрических оптимизационных моделях.

Ключевые слова: алгоритмы кластеризации, агропромышленный комплекс, GH-VNS, подход, предиктивное управление, АПК.

ON APPLICATION OF ONE APPROACH TO PROCESSING LARGE DIMENSIONAL DATA

Rozhnov Ivan Pavlovich, candidate of technical sciences,
senior researcher, Institute of Informatics and Telecommunications
Reshetnev Siberian State University of Science and Technology, Krasnoyarsk, Russia
e-mail: *ris2005@mail.ru*
Kazakovtsev Lev Aleksandrovich, doctor of technical sciences, associate professor,
professor of the department of “Information Technologies and Mathematical Support of Information Systems”, Institute of economics and agro-industrial complex management
Krasnoyarsk state agrarian university, Krasnoyarsk, Russia
Reshetnev Siberian State University of Science and Technology, Krasnoyarsk, Russia
e-mail: levk@bk.ru
Rezova Natalia Leonidovna,
associate professor of the department of “Information and Control Systems”, Institute of Informatics and Telecommunications
Reshetnev Siberian State University of Science and Technology, Krasnoyarsk, Russia
e-mail: natalyakl@yandex.ru

Abstract. For processing large data in the agro-industrial complex, the application of an approach to the development of automatic grouping algorithms based on parametric optimization models.

Key words: clustering algorithms, agro-industrial complex, GH-VNS, approach, predictive control, AIC.

В настоящее время возрастающее использование данных большого объема стимулирует разработку и применение методов и средств обработки и анализа массивных данных огромных объемов в различных областях и сферах жизнедеятельности, это и промышленность, энергетика, медицина, ГО и ЧС, логистика, государственное управление и т.д. Эти отрасли успешно развиваются вместе с прогрессом в сфере цифровых технологий. Но, в агропромышленном комплексе (АПК)

России, хоть он и является одной из важнейших отраслей, применяются далеко не самые новые методы производства сельхозпродукции и цифровые технологии распространены слабо [4, 9].

На современном этапе развития в агропромышленном комплексе происходит увеличение не только объема данных, но и рост количества и качества современных средств и решений, а следовательно, потребность в анализе и достоверных выводах для принятия управленческих решений [2, 4].

Процесс цифровизации в агропромышленном комплексе – это и возможность создавать сложные производственно-логистические цепочки, охватывающие сельхозпроизводителей и их поставщиков, логистику и оптово-розничные компании в единый комплекс с предиктивным управлением. Это позволит существенно снизить себестоимость сельхозпродукции, увеличив, таким образом, объемы производства и продаж, и доступность продуктов питания для потребителей [6, 9].

К примеру, используя методы искусственного интеллекта, можно оценить потенциальные недостатки питательных веществ в почве. Система отследит изменения в растениях, на которые влияют дефекты почвы и вредители, или болезни растений, которые распространяются по полю. Проанализировав проблему, сельхозпроизводителю система порекомендует методы восстановления почвы, а также и другие решения, повышающие качество и количество урожая. Однако конкретных результатов подобных исследований в АПК пока ничтожно мало. Требуется дополнительное изучение применения данных технологий, чтобы понять действенность и полезность для реального применения, в том числе и в качестве предиктивного управления.

Одним из перспективных направлений в аналитике больших данных (Big Data) является кластерный анализ, спектр использования которого очень широк и применяется для решения задач практически во всех сферах жизнедеятельности человека [1, 12].

Совместное применение метода жадных эвристик [13] с VNS-алгоритмами [11] для задач k-средних [10, 15], k-медоид [14] и алгоритма СЕМ [2, 3] было ранее подробно рассмотрено в работах [1, 2, 5]. Для увеличения точности вычислений алгоритмов автоматической группировки был применен подход к разработке алгоритмов кластеризации, основанных на параметрических оптимизационных моделях, с комбинированным применением алгоритмов поиска с чередующимися рандомизированными окрестностями и жадных агломеративных эвристических процедур [1, 7]. Общая схема данного подхода представлена на рисунке 1 [1].

Для нашего исследования мы использовали информацию из открытых данных Министерства сельского хозяйства Российской Федерации, таких как, например, «Каталог пестицидов, зарегистрированных на территории Российской Федерации».

При кластеризации наборов данных каждый из алгоритмов был запущен по 30 раз. Из результатов вычислительных экспериментов по каждому алгоритму рассчитаны значения целевой функции: минимальное значение (Min), максимальное значение (Max), среднее значение (Среднее) и среднеквадратичное отклонение (СКО). Алгоритмы k-means и j-means запускались в режиме мультистарта (таблица 1).

Таблица 1 –**Результаты вычислительных экспериментов**

Алгоритм	Значение целевой функции			
	Min	Max	Среднее	Среднеквадратичное отклонение
k-means	3 743,40	3 744,62	3 743,39	0,9346
j-means	3 742,07	3 743,52	3 742,57	0,4487
GH-VNS1	3 741,97	3 743,08	3 742,36	0,4020
GH-VNS2	3 741,97	3 743,15	3 742,06	0,5028
GH-VNS3	3 741,97	3 742,10	3 741,99	0,0424
ГАЖЭ+ЛП	3 742,10	3 745,73	3 743,72	1,2199
ГА ФП	3 741,99	3 742,34	3 742,10	0,2045
ГА классический	3 742,09	3 742,88	3 742,45	0,3489

В таблице 1 использованы следующие сокращения и аббревиатуры: GH-VNS - алгоритм кластеризации разработанный с помощью рассматриваемого подхода, ГА – генетический алгоритм, ГАЖЭ+ЛП – генетический алгоритм с жадной эвристикой с вещественным алфавитом и локальным поиском, ГА ФП – генетический алгоритм с рекомбинацией подножеств фиксированной длины.

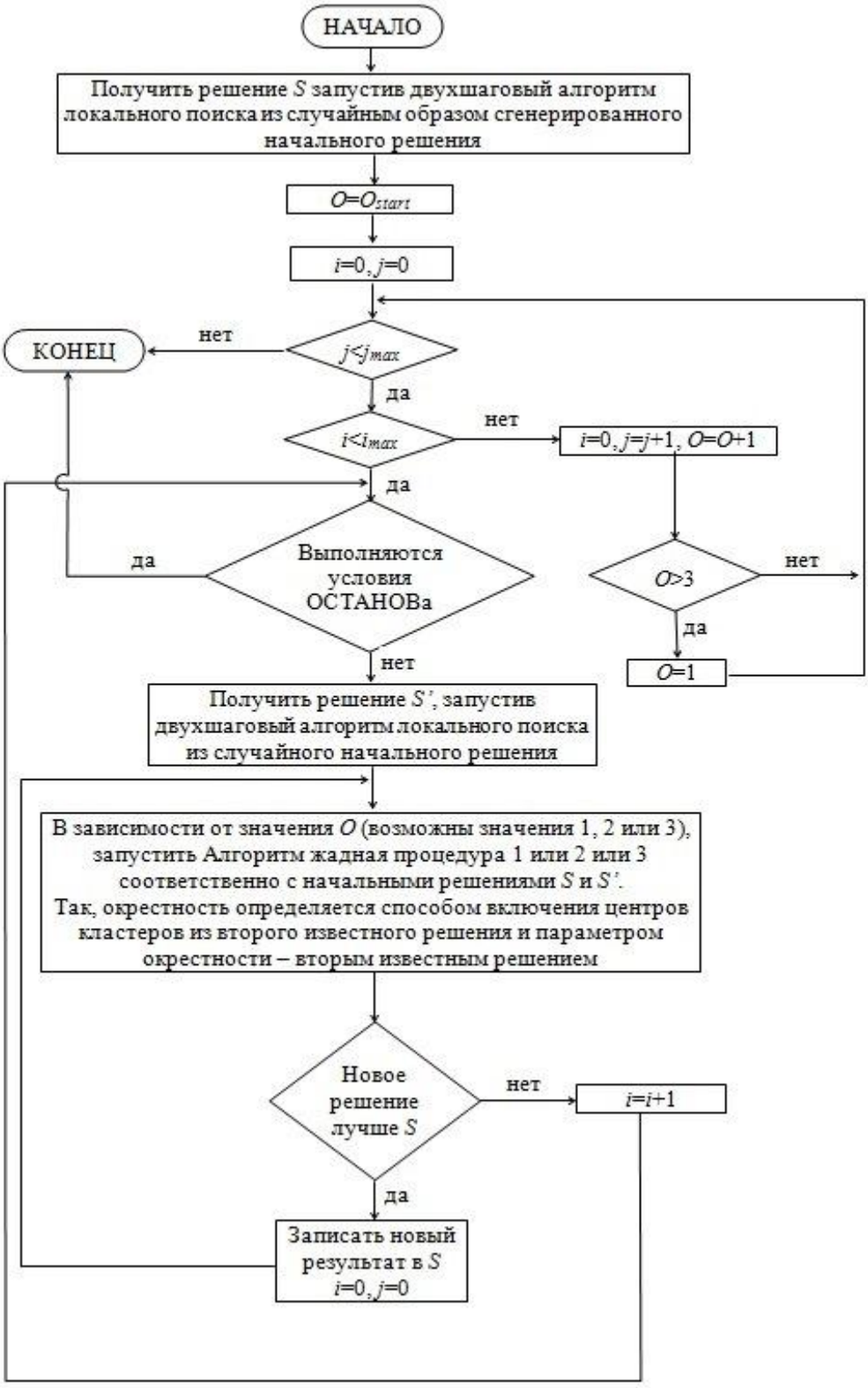

Рисунок 1 - Общая схема подхода к разработке алгоритмов кластеризации

По результатам вычислительных экспериментов видно, что алгоритмы GH-VNS разработанные с помощью представленного подхода (рисунок 1) имеют более точные (меньшее среднее значение целевой функции) и более стабильные (меньшее среднеквадратичное отклонение целевой функции) показатели в сравнении со считающимися классическими алгоритмами (k-means, j-means, PAM и CEM), а также некоторыми генетическими алгоритмами. Таким образом, рассматриваемый в данной статье подход можно применять для обработки больших данных в агропромышленном комплексе.

В то же время в условиях сложности автоматической группировки объектов для оценки качества алгоритмов кластеризации желательно использовать несколько способов оценки для получения более точных результатов, на основе которых можно проверить результат работы тех или иных алгоритмов кластеризации. Так как алгоритмы кластерного анализа не являются универсальными, то в настоящее время существует тенденция к применению коллективных (ансамблевых) подходов [8], что позволяет снизить зависимость конечного решения от заданных параметров исходных алгоритмов и получить более устойчивое и точное решение [16].

Список литературы

1. Алгоритмы автоматической группировки с повышенными требованиями к точности и стабильности результата / И. П. Рожнов, Л. А. Казаковцев, В. И. Орлов, Д. Л. Михнев ; Сибирский государственный университет науки и технологий им. акад. М.Ф. Решетнева. Москва : Издательский Дом "Инфра-М", 2020. 192 с. (Научная мысль). ISBN 9785160166414.
2. Казаковцев, Л. А. Усовершенствованный сем-алгоритм для данных большой размерности / Л. А. Казаковцев, И. П. Рожнов, П. Ф. Шестаков // Наука и образование: опыт, проблемы, перспективы развития : материалы международной научно-практической конференции, Красноярск, 16–18 апреля 2019 года / Красноярский государственный аграрный университет. – Красноярск: Красноярский государственный аграрный университет, 2019. С. 280-284.
3. Королёв В.Ю. ЕМ-алгоритм, его модификации и их применение к задаче разделения смесей вероятностных распределений. Теоретический обзор. ИПИ РАН. М.2007.с. 94.
4. Москалев, С. М. Искусственный интеллект и интернет вещей как инновационные методы совершенствования агропромышленного сектора / С. М. Москалев, Н. В. Клименок-Кудинова // Известия Санкт-Петербургского государственного аграрного университета. 2018. С. 121-130.
5. Рожнов, И. П. Алгоритмы с чередованием жадных эвристических процедур для дискретных задач кластеризации / И. П. Рожнов // Системы управления и информационные технологии. 2019. № 1(75). С. 49-55.
6. Рожнов, И. П. Анализ влияния тарифов в топливно-энергетическом комплексе на развитие региона в послереформенные годы / И. П. Рожнов, Л. А. Казаковцев // Проблемы современной аграрной науки: материалы международной заочной научной конференции, Красноярск, 15 октября 2014 года / Ответственные за выпуск: Г.И. Цугленок, Ж.Н. Шмелева. Красноярск: Красноярский государственный аграрный университет, 2015. С. 68-71.
7. Рожнов, И. П. Подход к разработке алгоритмов автоматической группировки на основе параметрических оптимизационных моделей / И. П. Рожнов, Л. А. Казаковцев // Информатика и системы управления. – 2020. – № 1(63). – С. 24-37. – DOI 10.22250/isu.2020.63.24-37.
8. Составление оптимальных ансамблей алгоритмов кластеризации / И. П. Рожнов, В. И. Орлов, М. Н. Гудыма, В. Л. Казаковцев // Системы управления и информационные технологии. 2018. № 2(72). С. 31-35.
9. Шувалов, А.А. Интернет вещей как инновационные методы совершенствования агропромышленного сектора / А. А. Шувалов // Вестник науки. 2019. Т. 1. № 7(15). С. 91-96.
10. Drezner Z. Facility Location: Applications and Teory / Z. Drezner, H. Hamacher // Springer-Verlag, Berlin, Germany, 2004.
11. Hansen P. Variable neighborhood search: principles and applications / Hansen P., Mladenovic N. // Eur. J. Oper. Res. 2001. Vol.130. P.449–467.
12. Jain, A.K. Data clustering: 50 years beyond K-means / A.K. Jain // Pattern Recognition Letters. 2010. Vol. 31. P. 651-666.
13. Kazakovtsev L.A. Genetic Algorithm with Fast Greedy Heuristic for Clustering and Locagtion Problems / Kazakovtsev L.A., Antamoshkin A.N. // Informatica (Ljubljana). 2014. Т. 38. № 3. С. 229-240.
14. Kazakovtsev, L. Self-configuring (1 + 1)-evolutionary algorithm for the continuous p-median problem with agglomerative mutation / L. Kazakovtsev, I. Rozhnov, G. Shkaberina // Algorithms. 2021. Vol. 14. No 5. DOI 10.3390/a14050130.
15. MacQueen, J.B. Some Methods of Classification and Analysis of Multivariate Observations. In Proceedings of the 5th Berkley Symposium on Mathematical Statistics and Probability, Berkeley, CA, USA, June 21–July 18, 1965 and December 27, 1965–January 7, 1966; *1*, pp. 281–297.
16. Rozhnov, I. Ensembles of clustering algorithms for problem of detection of homogeneous production batches of semiconductor devices / I. Rozhnov, V. Orlov, L. Kazakovtsev // CEUR Workshop Proceedings, Omsk, 08–14 июля 2018 года. Omsk, 2018. P. 338-348.