

На правах рукописи



Гудыма Михаил Николаевич

**АЛГОРИТМЫ РЕШЕНИЯ СЕРИИ ЗАДАЧ
АВТОМАТИЧЕСКОЙ ГРУППИРОВКИ**

Специальность: 05.13.01 – Системный анализ, управление и обработка информации (космические и информационные технологии)

АВТОРЕФЕРАТ

диссертации на соискание ученой степени
кандидата технических наук

Красноярск – 2017

Работа выполнена в ФГБОУ ВО "Сибирский государственный университет науки и технологий им. академика М.Ф.Решетнева", г.Красноярск.

Научный руководитель:

Казаковцев Лев Александрович
доктор технических наук, доцент

Официальные оппоненты:

Кочетов Юрий Андреевич
доктор физ.-мат. наук, профессор,
Институт математики им. С.Л.Соболева
СО РАН (г.Новосибирск),
главный научный сотрудник
лаборатории «Математические модели
принятия решений»

Крутиков Владимир Николаевич
доктор технических наук, доцент,
ФГБОУ ВО «Кемеровский
государственный университет»,
профессор кафедры прикладной
математики

Ведущая организация:

ФГБОУ ВО "Воронежский государственный
технический университет"

Защита состоится 8 декабря 2017 г. в 14:30 часов на заседании диссертационного совета Д 212.249.05, созданного на базе ФГБОУ ВО "Сибирский государственный университет науки и технологий им. академика М.Ф.Решетнева" по адресу: 660037, г.Красноярск, просп.им.газеты Красноярский рабочий, 31.

С диссертацией можно ознакомиться в библиотеке Сибирского государственного университета науки и технологий им. академика М.Ф.Решетнева.

Автореферат разослан “ ____ ” _____ 2017 г.

Ученый секретарь
диссертационного совета

Илья Александрович Панфилов

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность. Настоящая работа посвящена разработке и исследованию алгоритмов автоматической группировки (кластеризации), широко используемой в системах интеллектуального анализа данных. Кластеризация, основываясь на установленном отношении схожести элементов, устанавливает подмножества (кластеры), в которые группируются входные данные. Одними из простейших и эффективнейших являются методы и модели, основанные на минимизации суммарных расстояний между объектами одной группы (кластера) или между объектами кластера и его центром. Данные модели имеют сходство, а иногда идентичны моделям теории размещения, в частности – р-медианной модели. Задача поиска центра кластера при этом связана с решением классической задачи теории размещения – задачи Вебера. Задачи автоматической группировки данных могут накладывать специфические требования по учету расстояний в пространствах этих данных. Поэтому назрела необходимость расширить перечень используемых моделей и алгоритмов размещения с различными мерами расстояния. Иной подход представляют модели, основанные на разделении смеси распределений. Требуется определить, каким из распределений порожден тот или иной объект.

Все эти задачи NP-трудны, для их приближенного решения разработано большое количество, в основном, рандомизированных алгоритмов. В частности, алгоритмы метода жадных эвристик¹ позволяют получать результаты, превосходящие по точности (выраженной значением целевой функции) и стабильности (воспроизводимости) результаты других известных методов при объеме данных до сотен тысяч векторов данных большой размерности.

Большинство алгоритмов для данных задач требуют указания числа групп (кластеров). Определение этого числа является отдельной трудной задачей, решаемой с использованием различных критериев. Другим подходом является решение серии задач. Под серией понимается множество задач, различающихся только числом групп (кластеров), на которые разбиваются объекты. Результаты решения этих задач в дальнейшем могут быть проанализированы с использованием любых критериев. Примером такой задачи с неизвестным числом кластеров (групп) является задача разделения сборной партии электрорадиоизделий космического применения на однородные производственные партии, изготовленные из единой партии сырья по данным сотен неразрушающих тестовых испытаний в специализированных тестовых центрах. На выходе имеем массив результатов этих измерений, который, кроме основной задачи отсева некачественных изделий, может быть задействован для решения задачи выявления однородных партий изделий, число которых неизвестно. Данная задача в рамках производственного процесса тестирования должна быть решена сравнительно быстро, при этом результат должен быть таким, чтобы его трудно было улучшить любым известным методом без значительных вычислительных затрат.

Степень разработанности темы. Стоит отметить, что довольно долго теория размещения и кластерный анализ развивались параллельно, и их методы схожи. В 1909 г. А.Вебер исследовал задачу о нахождении центра тяжести для взвешенных точек (задача Вебера или 1-медианная задача), являющуюся развитием задачи

¹ Kazakovtsev L. A., Antamoshkin A. N. Greedy heuristic method for location problems // Вестник СибГАУ. 2015. №2 С.317-325.

Kazakovtsev L. A., Antamoshkin A. N. Genetic Algorithm with Fast Greedy Heuristic for Clustering and Location Problems // Informatica (Ljubljana), 2014, Vol.38 (3), pp.229–240.

П. Ферма XVII века. В 1937 г. А.Вайсфельд предложил решение этой задачи градиентным спуском. В середине прошлого века С.Л.Хаками рассматривал задачу нахождения медианы графа, и определил возможность дискретизации непрерывной задачи Вебера. Позднее Хаками обобщил эту задачу до нахождения p медиан графа с минимальной суммой взвешенных расстояний.

Одной из наиболее популярных моделей кластерного анализа, модель k -средних, была предложена в 1957 г. Г.Штейнгаузом, алгоритм разработан С.Ллойдом тогда же, однако его работа была опубликована только в 1982 г. Она сходна с p -медианной задачей не только по своей постановке, но и по используемым традиционным подходам к ней – ALA-процедура (Л.Купер, 1964) и процедура k -средних построены по одной схеме. Подобную связь можно заметить и между агломеративными эвристическими процедурами, применяемыми в кластерном анализе.

Среди ученых, в чьих трудах получили развитие теория размещения и автоматическая группировка объектов, в первую очередь необходимо упомянуть Ц.Дрезнера, С.Л.Хаками и Г.Весоловски. Весомый вклад внесли также В.А.Трубин, Н.Младенович, Дж.Бримберг и Р.Ф.Лав. Методы и модели теории размещения получили наиболее полное отражение в монографиях под редакцией Ц.Дрезнера и Х.Хамахера (2004), Р.Фарахани и М.Хекматфара (2009). В СССР исследованием вопроса размещения предприятий занимались В.Р.Хачатуров и В.П.Черенин. В Институте математики им. С.Л. Соболева СО РАН при разработке моделей стандартизации и унификации в качестве основы использовались модели размещения. Работы Э.Х.Гимади, В.Л.Береснева, А.А.Колоколова, а позже Ю.А.Кочетова, А.В.Еремеева, Г.Г.Забудского и др. заложили основу для разработки программно-математического аппарата решения этих задач.

В 2014 году Л.А.Казаковцевым и А.Н.Антамошкиным был предложен метод жадных эвристик. Метод широко использует эволюционные подходы, большой вклад в развитие которых вносит Красноярская школа эволюционных алгоритмов (Е.С.Семенкин и др.).

Настоящая работа посвящена разработке алгоритмов одновременного решения серий задач автоматической группировки объектов с различным числом групп (кластеров), позволяющих получать наиболее точные (по значению целевой функции) результаты по сравнению с другими известными методами.

Основная **идея** настоящей диссертации состоит в разработке генетического алгоритма метода жадных эвристик для задач автоматической группировки с заранее неизвестным числом групп (кластеров), который одновременно оперирует смешанной популяцией, состоящей из решений с различным числом групп (кластеров). Эффективность алгоритма повышается за счет того, что решения, являющиеся локальными оптимумами задач с различным числом групп (кластеров), совместно участвуют в едином процессе рекомбинации и создания новых решений.

Объектом диссертационного исследования являются задачи автоматической группировки многомерных данных с неизвестным числом групп (кластеров), **предметом исследования** – алгоритмы их решения.

Целью исследования является повышение эффективности систем автоматической группировки объектов, к которым предъявляются высокие требования по точности результата, выраженного значением целевой функции.

В процессе достижения поставленной цели решались следующие **задачи**:

1) анализ проблем, возникающих при применении методов кластеризации, основанных на моделях теории размещения и моделях разделения смеси распределений, при заранее неизвестном числе кластеров (групп);

2) разработка алгоритмов одновременного решения серии задач автоматической группировки (кластеризации) данных большой размерности (до сотен измерений) и большого объема (до десятков тысяч векторов данных) на основе моделей k -средних, k -медоид и k -медиан, различающихся только числом групп (кластеров), на которые разбиваются объекты, при заранее известном максимальном числе групп;

3) разработка алгоритма одновременного решения серии задач нечеткой кластеризации данных большой размерности на основе модели разделения смеси вероятностных распределений, различающихся только числом групп (кластеров, распределений), на которые разбиваются объекты, при заранее известном максимальном числе распределений;

4) разработка эффективных алгоритмов решения задач Вебера (1-медианных) с нестандартными мерами расстояния, востребованными в при решении практических p -медианных задач.

Методы исследования. Методологической базой явились работы по методам кластеризации, в частности – по методу жадных эвристик для задач автоматической группировки объектов. Использован математический аппарат теории размещения, методы теории оптимизации, в частности – эволюционные методы оптимизации, а также методы системного анализа, исследования операций, аналитической геометрии и теории вероятностей.

Новые научные результаты и положения, выносимые на защиту:

1. Предложен новый генетический алгоритм с вещественным алфавитом для одновременного решения серий непрерывных и дискретных задач автоматической группировки объектов на основе моделей k -средних, k -медоид, k -медиан с различными мерами расстояния, основанный на комбинации особой модификации жадных агломеративных эвристических процедур, эволюционных методов оптимизации, методов локального поиска и методов решения задачи поиска центра группы (задачи Вебера с соответствующей мерой расстояния). Разработанный алгоритм обеспечивает лучшие (по значению целевой функции) результаты для серии задач с числом групп (кластеров) от 2 до заданного значения p_{max} за приемлемое время для задач с большим объемом входных данных в сравнении с известными методами. Алгоритм позволяет одновременно решать серию задач, благодаря чему повышается эффективность систем автоматической группировки объектов.

2. Впервые предложен генетический алгоритм с вещественным алфавитом для одновременного решения серий задач автоматической группировки объектов на основе моделей разделения смеси вероятностных распределений, основанный на комбинации особой модификации жадных агломеративных эвристических процедур и ЕМ-алгоритма. Разработанный алгоритм обеспечивает результаты для серии задач с числом групп (кластеров) от 2 до заданного значения p_{max} , не уступающие результатам известных методов, позволяя, в отличие от них, одновременно решать за приемлемое время серию задач с большим объемом входных данных, за счет чего повышается эффективность работы систем автоматической группировки объектов.

3. Предложен новый алгоритм решения задачи Вебера с допустимыми зонами, ограниченными окружностями, который является составной частью алгоритма

решения задачи автоматической группировки с особой мерой расстояния, ограниченной снизу. Экспериментально показано, что алгоритм позволяет получать более точные решения в сравнении с известными алгоритмами.

Значение для теории. Результаты исследования дополняют арсенал эффективных эвристических методов решения NP-трудных задач автоматической группировки и размещения с широким кругом используемых мер расстояний. Принцип скрещивания в единой популяции генетического алгоритма особей, представляющих собой решения различных задач, различающихся единственным параметром – числом групп, создает основу для синтеза новых эффективных алгоритмов.

Практическая ценность методов решения задач автоматической группировки и задач размещения обусловлена широким диапазоном их применения как в задачах кластерного анализа или автоматической группировки данных, так и непосредственно в практических задачах группировки физических объектов или оптимального размещения в пространстве. Разработанный метод позволяет повысить эффективность алгоритмов за счет одновременного решения за ограниченное время сразу серии задач с большим объемом входных данных.

Практическая реализация результатов: Программная реализация алгоритма решения серии задач автоматической группировки с использованием модели k-средних с прямоугольной метрикой позволила повысить эффективность СППР автоматической классификации электрорадиоизделий по производственным партиям ОАО ИТЦ – НПО ПМ (г.Железногорск) за счет возможности работы с данной СППР в интерактивном режиме и встроить классификацию электрорадиоизделий в производственный процесс проведения испытаний.

Апробация работы. Основные положения и результаты диссертационной работы докладывались и обсуждались на международных конференциях. В их числе: XIII Международная научно-практическая конференция «Перспективы развития информационных технологий» (2013 г., г.Новосибирск), XV Международная научно-практическая конференция «Актуальные вопросы науки» (2014 г., г.Москва), XVIII Международная научная конференция «Решетневские чтения» (2014 г., г.Красноярск), Международная научная конференция «Проблемы современной аграрной науки» (2013 г., г.Красноярск), 6th International Congress on Ultra Modern Telecommunications and Control Systems and Workshops, (2014 г., г. Санкт-Петербург), XX Международная научная конференция «Решетневские чтения» (2016 г., г.Красноярск). Работа в целом обсуждалась на научно-техническом семинаре «Электронная компонентная база космических систем» (2016 г., г.Железногорск).

Публикации. Основные теоретические и практические результаты диссертации опубликованы в 17 статьях (также имеются 3 свидетельства о государственной регистрации программ для ЭВМ), среди которых 7 работ в ведущих рецензируемых изданиях, рекомендуемых в действующем перечне ВАК, 2 – в международных изданиях, проиндексированных в системах цитирования Scopus.

Структура и объем диссертации. Диссертация состоит из введения, 3 глав и заключения. Она изложена на 201 листе машинописного текста, содержит список литературы из 289 наименований.

ОСНОВНОЕ СОДЕРЖАНИЕ РАБОТЫ

Во **введении** обоснована актуальность, научная новизна и практическая значимость работы, изложены цели, задачи и методы исследования.

Глава 1 носит обзорный характер. Делается обзор постановок задач автоматической группировки и методов их решения.

Среди методов интеллектуального анализа данных особое место занимают классификация и автоматическая группировка (кластеризация). Классификация, по известной заранее группировке данных на подмножества (классы), называемой обучающей выборкой, устанавливает закономерность, по которой данные группируются именно таким образом. Кластеризация, основываясь на установленной мере схожести объектов, устанавливает группы объектов (кластеры), в которые группируются входные данные. В методах нечеткой кластеризации объекты входного множества относят к тому или иному кластеру на основании значения функции принадлежности. Примером такой функции принадлежности является функция вероятностного распределения. В этом случае задача сводится к отысканию параметров распределений, которыми могут быть порождены объекты, а в качестве целевой функции используется логарифмическая функция правдоподобия. Нечеткая кластеризация является одной из наиболее проработанных методик интеллектуального анализа данных, в то же время ее методы достаточно громоздки в вычислительном плане.

Задача k -средних может быть классифицирована как задача теории размещения: целью является нахождение k точек (центров, центроидов) $X_1, \dots, X_k$ в d -мерном пространстве, таких, чтобы сумма квадратов расстояний от известных точек (векторов данных) $A_1, \dots, A_N$ до ближайшей из искомых точек достигала минимума:

$$\arg \min F(X_1, \dots, X_k) = \sum_{i=1}^N \min_{j \in \{1, k\}} \|X_j - A_i\|^2.$$

Цель непрерывной p -медианной (k -медианной) задачи – нахождение k точек, таких, чтобы сумма расстояний (в какой-либо метрике) от известных точек (векторов данных) до ближайшей из искомых точек достигала минимума:

$$\arg \min F(X_1, \dots, X_k) = \sum_{i=1}^N \min_{j \in \{1, k\}} L(X_j, A_i). \quad (1)$$

Здесь $L()$ – некая функция расстояния между двумя точками в d -мерном пространстве.

Непрерывные задачи размещения, включая p -медианную, хорошо изучены с такими метриками, как евклидова, манхэттенская (прямоугольная), чебышёвская и другие варианты норм Минковского. Многие исследования посвящены решению 1-медианной задачи (задачи Вебера) с этими метриками, которая является составной частью наиболее известных алгоритмов для p -медианной задачи. В то же время, практические задачи могут требовать использования специфических мер расстояния. Так, например, одним из условий задачи может являться обязательное пренебрежение расстояниями меньше некоторого минимального предела.

Традиционно собственно p -медианной задачей называют задачу с евклидовой метрикой $L(X_j, A_i) = \sqrt{\sum_{k=1}^d (x_{j,k} - a_{i,k})^2}$. Здесь $X_j = (x_{j,1}, \dots, x_{j,d})$, $\forall j = 1, p$, $A_i = (a_{i,1}, \dots, a_{i,d})$, $\forall i = 1, N$. В случае, если в качестве меры расстояния используется квадрат евклидовой метрики $L(X_j, A_i) = \sum_{k=1}^d (x_{j,k} - a_{i,k})^2$ имеем классическую задачу k -средних.

Для решения задачи k -средних и p -медианной задачи может быть использован

одноименный алгоритм, называемый также ALA-алгоритмом (Alternating Location-Allocation – чередующееся размещение-распределение), включающий два чередующихся шага:

Алгоритм 1. ALA-процедура. Дано: векторы данных $A_1...A_N$, k начальных центров кластеров $X_1...X_k$.

1. Для каждого центра X_i составить кластер (множество) C_i векторов данных, для которых этот центр является ближайшим.

2. Для каждого кластера C_i рассчитать новое значение центра X_i (т.е. решить задачу Вебера).

3. Повторить с шага 1, если шаги 1 и 2 привели к каким-либо изменениям.

За исключением особых случаев, задачи k -средних и p -медианная являются NP-трудными, требующими глобального поиска. Результат ALA-процедуры зависит от выбора начальных центров кластеров и это является проблемой с точки зрения обеспечения воспроизводимости результатов вычислений (результаты разбиения на кластеры могут сильно различаться).

Генетические алгоритмы (ГА) с жадной агломеративной эвристикой, изначально разработанные для p -медианной задачи на сети О.Алпом, Э.Эркутом и Ц.Дрезнером, в дальнейшем адаптированные для непрерывных задач и включенные в состав метода жадных эвристик Л.А.Казаковцева, являются компромиссными по точности результата и вычислительным затратам. Эти алгоритмы и новые алгоритмы, предложенные в диссертации, являются ГА с вещественным алфавитом (“особи” кодируются не L -битной строкой).

Алгоритм 2. ГА с жадной эвристикой для p -медианной задачи (с полным объединенным решением). Дано: Размер популяции N_p , число кластеров p .

1. Сформировать (случайным образом или с использованием процедуры k -means++) N_p различных начальных решений $\chi_1, ..., \chi_{N_p} : \chi_i \subset \mathbb{R}^d, |\chi_i| = p \forall i = 1, \overline{N_p}$ – множеств индексов векторов данных мощности p , используемых в качестве исходных решений ALA-алгоритма. Для каждого из начальных решений оценить значение целевой функции $F_{fitness}(\chi)$, сохранить значения данной функции в переменных $f_1, ..., f_{N_p}$.

2. Если достигнуты условия останова, то ОСТАНОВ. Решением является начальное решение χ_{i^*} , которому соответствует наименьшее значение f_i . Для нахождения окончательного решения запускается ALA-алгоритм (Алгоритм 1).

3. Выбрать случайным образом два индекса $k_1, k_2 \in \{1, \overline{N_p}\}, k_1 \neq k_2$.

4. Получить промежуточное решение $\chi_c = \chi_{k_1} \cup \chi_{k_2}$.

5. Если $|\chi_c| \leq p$, то перейти к Шагу 7:

6. Вычислить $j^* = \arg \min_{j \in \chi_c} F_{fitness}(\chi_c \setminus \{j\})$. Исключить j^* из χ_c : $\chi_c = \chi_c \setminus \{j^*\}$.

Запустить ALA-алгоритм для решения χ_c и перейти к Шагу 5.

7. Если $\exists i \in \{1, \overline{N_p}\} : \chi_i = \chi_c$, то перейти к Шагу 2.

8. Выбрать индекс $k_3 \in \{1, \overline{N_p}\}$ с помощью турнирного замещения: выбираются случайным образом два индекса $k_4, k_5 \in \{1, \overline{N_p}\}$, если $f_{k_4} > f_{k_5}$, то $k_3 = k_4$, иначе $k_3 = k_5$; заменить χ_{k_3} и соответствующее значение целевой функции: $\chi_{k_3} = \chi_c$,

$f_{k_3} = F_{fitness}(\chi_c)$. Перейти к Шагу 2.

Дальнейшим развитием стал алгоритм с частичным объединенным решением, дающий лучшие результаты при большом числе кластеров ($p > 30$). Шаг 4 получил следующую редакцию:

4. С помощью генератора случайных чисел выбрать $r_{init} \in [0; 1]$; $r = 1 + [(p-1) \cdot r_{init}^2]$; из множества χ_{k_2} выбрать случайным образом подмножество $\chi_{k_2}^*$ мощности r ; Получить промежуточное решение $\chi_c = \chi_{k_1} \cup \chi_{k_2}^*$, случайным образом выбрать подмножество.

Большинство алгоритмов для решения задачи k-средних и p-медианной задачи, таких как ALA-процедура или весьма эффективный ГА с рекомбинацией подмножеств фиксированной длины В.Шена и С.Лиу требуют, чтобы число кластеров p было известно. Другие алгоритмы, как например X-means, выбирают наилучшее значение p в соответствии со специальным критерием, при этом точность решения весьма невысока. Выбор адекватного критерия – отдельная сложная задача.

К типичным задачам, где значение числа групп объектов заранее неизвестно, относится задача разделения сборной партии электрорадиоизделий космического применения на однородные производственные партии. Производители электронных компонентов в США и ЕС выпускают специальный класс продукции для космической отрасли (Space). В России такое специализированное производство отсутствует, чем и обусловлена необходимость комплексных испытаний продукции в специализированных тестовых центрах. Характеристики каждого изделия из партии проверяется сотнями неразрушающих тестов. Выборочно применяется разрушающий физический анализ. Результаты разрушающих испытаний нельзя распространять на всю партию, если партия неоднородна. Для исследования партии на однородность достаточно данных неразрушающих испытаний. Для разбиения поступившей (предположительно сборной) партии изделий на несколько предполагаемых однородных производственных партий может эффективно применяться метод кластерного анализа k-средних с прямоугольной метрикой по данным неразрушающих тестовых испытаний, подвергнутым особому способу нормировки. При этом число однородных партий в сборной партии неизвестно и решается сразу серия задач автоматической группировки с различным числом предполагаемых однородных партий от 2 до 20. При этом результаты расчетов должны быть достаточно точны и воспроизводимы при многократных запусках алгоритма, что определяется необходимостью обеспечения возможности их проверки.

Широко известный ЕМ-алгоритм для разделения смеси k сферических или некоррелированных гауссовых распределений может быть описан следующим образом. Дан набор данных $S \subset R^n$. Алгоритм стартует с начальными стартовыми значениями параметров распределений $\mu_i^{(0)}, \alpha_i^{(0)}, \sigma_i^{(0)}$.

Алгоритм 3. ЕМ-алгоритм. t – номер итерации

Шаг 1. Пусть $\tau_i \sim N(\mu_i^{(t)}, \sigma_i^{(t)2} I_n)$ является плотностью i -го гауссова распределения:

$$\tau_i(x) = \frac{1}{(2\pi)^{n/2} \sigma_i^n} \exp\left(-\frac{\|x - \mu_i\|^2}{2\sigma_i^2}\right). \text{ Для каждого вектора исходных данных } x \in S \text{ и}$$

каждого $1 \leq i \leq k$ по формуле Байеса вычислим условную вероятность того, что x относится к i -му распределению с учетом текущих параметров:

$$p_i^{(t+1)}(x) = \alpha_i^{(t)} \tau_i(x) / \left(\sum_j \alpha_j^{(t)} \tau_j(x)\right).$$

Шаг 2. Производится адаптация параметров распределений. Пусть N – количество векторов данных. $\alpha_i^{(t+1)} = \frac{1}{N} \sum_{x \in S} p_i^{(t+1)}(x)$; $\mu_i^{(t+1)} = \frac{\sum_{x \in S} x p_i^{(t+1)}(x)}{N \alpha_i^{(t+1)}}$;

$$\sigma_j^{(t+1)2} = \sum_{i=1}^k \sum_{x \in S} \|x - \mu_i^{(t+1)}\|^2 p_i^{(t+1)}(x) / (N \alpha_i)$$

Повторять с шага 1.

Условием останова в таком алгоритме является останов приращенной целевой логарифмической функции правдоподобия $L^{(t+1)} = \sum_{x \in S} \sum_{i=1}^k \ln(\tau_i p_i^{(t+1)}(x))$.

Структура этого алгоритма повторяет структуру ALA-алгоритма. При этом для задач разделения смеси распределений число распределений, как правило, заранее неизвестно, и для решения этой проблемы могут быть применены те же подходы, что и к задачам k -средних, k -медоид, k -медиан с неизвестным числом кластеров.

Анализ известных задач, моделей и методов выявил, что методы, предполагающие одновременное с решением задачи автоматической группировки решение задачи определения числа групп в данных, обладают достаточно низкой точностью, кроме того, критерии оценки числа групп (кластеров) могут давать неточный результат. В связи с этим перспективной является разработка методов одновременного решения серии задач, различающихся только числом кластеров. Общность подходов и алгоритмов решения задач автоматической группировки, основанных на моделях теории размещения, с методами, основанных на разделении смеси вероятностных распределений, дает обоснованную надежду на применимость общих подходов к повышению эффективности алгоритмов.

Глава 2 посвящена непосредственно разработке новых алгоритмов решения серий задач автоматической группировки. В качестве простейшего подхода мы предлагаем минимальную модификацию жадной эвристики: после достижения требуемого числа кластеров p процесс исключения кластеров продолжается, и алгоритм продолжает фиксировать наилучшие известные значения целевой функции для каждого числа кластеров вплоть до 2 кластеров. Таким образом, мы можем получить решения серии задач с $p = 2, p_{\max}$. При этом максимальное предполагаемое значение числа кластеров $p_{\max}$ все же должно быть известно.

Алгоритм 4. ГА с жадной эвристикой для серии задач с $p \in \{2, p_{\max}\}$.

1. Инициализация популяции из N_{pop} особей. Каждая особь является множеством из $p_{\max}$ центров (обозначим их X_i). Присвоить $F_{new,j} = +\infty$ для каждого $j \in \{1, N_{pop}\}$ Инициализировать массивы значений целевой функции $F_k^* = +\infty$ и лучших решений $X_k^* = \{\}$ для каждого $k \in \{2, p_{\max}\}$.

2. Выбрать случайным образом $j_1, j_2 \in [1, N_{pop}]$, $j_1 \neq j_2$

3. $X_{new} = X_{j_1} \cup X_{j_2}$; //(получение объединенного решения)

4. Пока $|X_{new}| > p_{\max}$:

4.1. Выбрать элемент j , такой, чтобы его исключение давало наименьший прирост целевой функции: $j = \arg \min_{j \in X_{new}} F(X_{new} \setminus \{j\})$

4.2. $X_{new} = X_{new} \setminus \{j\}$. Следующая итерация 4.

5. Присвоить $F_{new} = 0$; $X^* = X_{new}$.

6. Пока $|X_{new}| > 2$:

6.1. Присвоить $F_{new} = F_{new} + f(X_{new})$; $k = |X_{new}|$; $F_k = f(X_{new})$; если $F_k < F_k^*$, то присвоить $F_k^* = F_k$;

6.2. Выполнить шаги 4.1 и 4.2 для X_{new} . Следующая итерация цикла 6.

7. Выбрать j_3 с использованием турнирного замещения по значению $F_{new,j}$. Присвоить $F_{j_3} = F_{new}$; $X_{j_3} = X^*$, $F_{new,j_3} = F_{new}$.

8. Проверить условия останова, перейти к шагу 2.

Как показывают вычислительные эксперименты, такой алгоритм дает хорошие решения с числом кластеров p , близким к максимальному p_{max} . При $p < p_{max}/2$ точность быстро падает в сравнении с известными методами решения единственной задачи.

Важным параметром эволюционных алгоритмов, включая генетические алгоритмы с жадной эвристикой, является размер популяции. В изначальном варианте генетического алгоритма с жадной эвристикой для p -медианной задачи предлагается такой размер популяции, чтобы каждый объект являлся центром хотя бы одного кластера хотя бы в одном из решений-«особей». Такой подход приводит к формированию больших популяций с учетом того, что жадная эвристика включает в себя запуск алгоритма локального поиска. В то же время, для очень больших задач время выполнения единственной процедуры «скрещивания» (шаги 3-6 Алгоритма 4) может быть велико и трудно предсказуемо, и в этой связи для очень больших задач также работа не доходит даже до третьего поколения «особей». Для малых же задач такая популяция быстро вырождается – алгоритм раз за разом генерирует одни и те же решения из очень ограниченного множества возможных комбинаций.

Мы предлагаем простую модификацию алгоритма с размером популяции, монотонно растущим с номером итерации.

Алгоритм 5. ГА с жадной эвристикой и динамической популяцией для решения серии задач с $p \in \{2, p_{max}\}$.

1. Инициализация начальной популяции начального размера из $N_{порнач}$ особей. Каждая особь является множеством из p_{max} центров (обозначим их X_i). Присвоить $F_{new,j} = +\infty$ для каждого $j \in \{1, N_{порнач}\}$. Инициализировать массивы значений целевой функции $F_k^* = +\infty$ и лучших решений $X_k^* = \{\}$ для каждого $k \in \{2, p_{max}\}$. $N_{iter} = 0$.

2. $N_{iter} = N_{iter} + 1$; $N_{пор} = \max\{N_{порнач}; \lceil \sqrt{1 + N_{iter}} \rceil + 2\}$; Если $N_{пор}$ изменилось, то инициализировать особь $X_{N_{пор}}$ аналогично шагу 1. Выбрать случайным образом $j_1, j_2 \in \{1, N_{пор}\}$, $j_1 \neq j_2$;

Дальнейшие шаги алгоритма повторяют Алгоритм 4 с шага 3.

Как показывают эксперименты, такой подход позволяет не задумываться об оптимальном размере популяции. Но он никак не решает проблему низкой точности результатов при $p < p_{max}/2$.

Исходный Алгоритм 4, как и Алгоритм 5 для решения серии задач, оперируют популяцией решений («особей»), каждое из которых состоит из одинакового числа p_{max} элементов – центров (центроидов, медоидов) кластеров. Для решения проблемы точности результата мы предлагаем модификацию алгоритма с гетерогенной популяцией – то есть популяцией, содержащей в своем составе решения («особи») различным числом центров p .

Идея заимствована из нашего алгоритма для решения задачи псевдобулевой оптимизации о загрузке производственных мощностей литейно-прокатного

производства – задаче многокритериальной оптимизации, решаемой алгоритмом метода жадных эвристик. Особи в популяции ГА – множества различной мощности.

Эффективность алгоритмов метода жадных эвристик с частичным объединенным решением для некоторых задач обусловлена тем, что в этих алгоритмах в промежуточном решении число кластеров меньше, чем при полном объединении двух родительских особей. Развитием этой идеи является скрещивание в единой популяции решений с различным числом кластеров.

Следующий алгоритм оперирует популяцией с различными значениями p .

Алгоритм 6. ГА с гетерогенной популяцией для p -медианной задачи.

1. Инициализация начальной популяции начального размера из $N_{popнач}$ особей. Каждая особь является множеством из p_{max} центров (обозначим их X_i). Присвоить $F_{new,j} = +\infty$ для каждого $j \in \{1, \overline{N_{popнач}}\}$. Инициализировать массивы значений целевой функции $F_k^* = +\infty$ и лучших решений $X_k^* = \{\}$ для каждого $k \in \{2, \overline{p_{max}}\}$. $N_{iter} = 0$.

2. $N_{iter} = N_{iter} + 1$; $N_{pop} = \max\{N_{popнач}; \lceil \sqrt{1 + N_{iter}} \rceil + 2\}$; Если N_{pop} изменилось, то инициализировать особь X_{Npop} аналогично шагу 1. Выбрать случайным образом $j_1, j_2 \in \{1, \overline{N_{pop}}\}$, $j_1 \neq j_2$;

3. ; $X_{new} = X_{j_1} \cup X_{j_2}$

4. Пока $|X_{new}| > p_{max}$;

4.1. Выбрать элемент j , такой, чтобы его исключение давало наименьший прирост целевой функции: $j = \arg \min_{j \in X_{new}} F(X_{new} \setminus \{j\})$

4.2 $X_{new} = X_{new} \setminus \{j\}$ Следующая итерация 4.

5. Выбрать случайным образом $p_{child} \in \{2, \overline{p_{max}}\}$. Если $p_{child} > |X_{new}|$, то $p_{child} = |X_{new}|$;

6. $f_{child, |X_{new}|} = f(X_{new})$;

7. Пока $|X_{new}| > p_{child}$

7.1. Выбрать элемент j , такой, чтобы его исключение давало наименьший прирост целевой функции: $j = \arg \min_{j \in X_{new}} f_{child}(X_{new} \setminus \{j\})$

7.2. $X_{new} = X_{new} \setminus \{j\}$. Следующая итерация 7.

8. Пока $|X_{new}| > 2$;

8.1. Присвоить $f_{child, |X_{new}|} = F(X_{new})$; $k = |X_{new}|$; $f_{k, |X_{new}|} = f(X_{new})$; если

$f_{k, |X_{new}|} < F_k^*$, то присвоить $F_k^* = f_{k, |X_{new}|}$;

8.2. Выполнить шаги 4.1 и 4.2 для X_{new} . Следующая итерация 8.

9. Выбрать $j_3 \in \{1, \overline{N_{pop}}\}$ с использованием турнирного замещения (Алгоритм 7).

Присвоить $X_{j_3} = X_{new}$; $f_{j_3,k} = f_{child}$

10. Проверить условия останова, перейти к шагу 2.

Алгоритм 7. Турнирное замещение для Алгоритма 6.

1. Выбрать случайно $j_1 \in \{1, \overline{N_{pop}}\}$, $j_2 \in \{1, \overline{N_{pop}}\}$, $j_1 \neq j_2$; Выбрать случайно $p_{compar} \in \{2, \min\{|X_{j_1}|, |X_{j_2}|\}\}$; если $f_{j_1, p_{compar}} < f_{j_2, p_{compar}}$ то $j_3 = j_2$, иначе $j_3 = j_1$.

Нашу задачу автоматической группировки можно рассматривать и как двухкритериальную: первый критерий – функция (1), второй – число кластеров.

В целях сравнения новые и известные алгоритмы запускались по 30 раз. Классические алгоритмы для решения единственной задачи запускались со всеми возможными вариантами числа кластеров от 2 до p_{max} .

В Таблице 1 приведены результаты работы алгоритмов с полным объединенным (АПОР) и частичным объединенным решением (АЧОР) для разных размеров популяции по данным отбраковочных испытаний ИС 2Д522Б ($N=3711$, $d=10$, метрика l_1 , $p=10$, $p_{max}=20$, $t=1$ мин), испытаний ИС 140УД25 ($N=523$, $d=42$, метрика l_1 , $p=10$, $p_{max}=20$, $t=1$ мин) и классического набора данных MissAmerica ($N=6480$, $d=16$, метрика l_2^2 , $p=75$, $p_{max}=100$, $t=2,5$ мин). В Таблице 2 приведены результаты работы предложенного алгоритма с гетерогенной популяцией в сравнении с лучшими результатами работы алгоритмов с полным объединенным и частичным объединенным решением и результатами алгоритма с динамической популяцией. Для сравнения в таблицу включены результаты полученные мультистартом процедуры k -средних (ALA-процедуры) и генетическим алгоритмом с рекомбинацией подмножеств фиксированной длины, результаты испытаний ИС 1526ТЛ1 ($N=1234$, $d=120$, специальная мера расстояния, $L(X,Y)=\max\{\|X-Y\|,1\}$, описанная в Главе 3, $p=10$, $p_{max}=10$, $t=2$ мин), а также набор данных Ionosphere из репозитория UCI ($N=351$, $d=34$, метрика l_2^2 , $p=3$, $p_{max}=10$, $t=40$ сек).

Таблица 1. Результаты работы алгоритмов с полным и частичным объединенным решением для разных размеров популяции. ($F(x)$ значение целевой функции, σ – стандартное отклонение)

Разм. попул	Пар-р	2Д522Б (АПОР)	140УД25 (АПОР)	MissAmer. (АПОР)	2Д522Б (АЧОР)	140УД25 (АЧОР)	MissAmer. (АЧОР)
5	$F(x)$	11778,05	980,67	755822,05	11777,99	979,37	755845,47
	σ	0,134	4,757	1597,645	0,127	5,030	591,256
10	$F(x)$	11778,02	976,64	755506,27	11778,01	976,77	755929,18
	σ	0,137	2,641	995,725	0,134	3,546	1466,276
20	$F(x)$	11777,97	975,29	756066,50	11777,95	976,15	756506,73
	σ	0,117	1,419	898,331	0,168	2,453	519,380
50	$F(x)$	11778,00	974,81	756960,20	11777,93	974,76	756969,27
	σ	0,119	0,239	1155,526	0,080	0,283	825,028
100	$F(x)$	11791,46	984,98	758777,26	11789,00	989,38	759518,74
	σ	17,304	6,496	1897,579	13,680	8,012	1601,17
best	$F(x)$	11777,97	974,81	755506,27	11777,93	974,76	755845,47
	σ	0,117	0,239	995,725	0,080	0,283	591,256

Как видно из таблиц, новый алгоритм с динамическим размером популяции (Алгоритм 5) в большинстве случаев не уступает лучшему из вариантов решения алгоритмов со статическим размером популяции.

Алгоритм 6 с гетерогенной популяцией, в отличие от серийного Алгоритма 5, с уменьшением числа кластеров p дает решения, не уступающие решениям алгоритма для решения единственной задачи (Алгоритм 2).

Общность ЕМ-алгоритма и АЛА-алгоритма позволяет применить общие подходы к повышению точности результатов. В работе на основе эволюционного (генетического) алгоритма Л.А.Казаковцева и Д.В.Сташкова для решения задачи разделения сферических и некоррелированных гауссовых распределений с гетерогенной популяцией, аналогичный Алгоритму 6.

Таблица 2. Сравнение результатов работы алгоритмов. АПОР – лучшие результаты алгоритма с полным объединенным решением, АЧОР – лучшие результаты алгоритма с частичным объединенным решением, АДП – алгоритм с динамической популяцией, АГП – алгоритм с гетерогенной популяцией, АЛА – АЛА-алгоритм, АРПФД – генетический алгоритм с рекомбинацией подмножеств фиксированной длины. ($F(x)$ значение целевой функции, σ – стандартное отклонение)

Задача	П-тр	АПОР	АЧОР	АДП	АГП	АЛА	АРПФД
2Д522Б, $p=10$	$F(x)$	11777,97	11777,93	11778,04	11776,48	11777,89	11775,56
	σ	0,117	0,080	0,131	1,879	0,023	2,559
2Д522Б, $p=15$	$F(x)$	9452,69	9449,48	9453,94	9456,17	9456,83	9454,98
	σ	3,395	3,870	1,958	2,621	4,817	4,817
140УД25, $p=10$	$F(x)$	974,81	974,76	974,77	974,51	974,76	977,11
	σ	0,239	0,283	0,283	0,370	0,283	5,461
140УД25, $p=20$	$F(x)$	660,17	659,91	659,84	661,23	667,68	663,32
	σ	1,016	0,416	0,243	1,081	1,188	0,763
MissAmer $p=75$	$F(x)$	755506,3	755845,47	756019,80	755422,09	765261,8	761278,34
	σ	995,725	591,256	566,549	1502,533	502,817	1755,331
MissAmer $p=50$	$F(x)$	827604,5	827866,55	827904,01	828391,37	833266,8	831822,57
	σ	925,538	621,880	366,072	717,021	1434,037	792,091
1526ТЛ1, $p=10$	$F(x)$	3440,06	3440,15	3442,64	3442,62	3454,09	3448,64
	σ	1,849	1,569	3,881	2,492	17,296	5,360
Ion, $p=3$	$F(x)$	8453,95	8452,32	8454,93	8450,95	8451,55	8451,54
	σ	4,533	3,292	4,747	0	2,444	2,447

Таблица 3. Сравнительные результаты серийного алгоритма решения задач разделения смеси распределений

Набор данных, число вектор., размерн.	Число класт. k , тип распр., время	Алгоритм	Ср. рез-т (лог. ф-ция пр.- подоб.)	Среднеквадр. откл. результатов
Europe (UCI), $N=169308$, $d=2$.	40, сфер, 1.5 часа	Новый	-3625694,1*	20,148
		VNS2	-3625691,7*	14,769
		VNS3	-3625748,7	15,402
		ЕМ	-3625957,3	49,561
		СЕМ	-3625779,0	25,064
		SEM	-3625740,2	29,064
Тесты ИС 1526ТЛ1, $N=1234$, $d=120$	5, сфер., 5 сек.	Новый	3673,671*	44,043
		VNS2	3543,974	144,691
		VNS3	3591,539	139,988
		ЕМ	3598,160	32,160

Если в задачах k -средних, k -медиан, k -медоид параметрами решения задачи являются исключительно множества координат центров кластеров $X_i=(x_{i,1}, \dots, x_{i,d})$, то в задаче о разделении смеси распределений решениями являются множества параметров распределений, дополненные значениями их весовых коэффициентов – априорных вероятностей распределений $\alpha_i^{(b)}$, т.е. в этом случае каждое решение X – это множество троек $(\mu_i^{(b)}, \sigma_i^{(b)}, \alpha_i^{(b)})$. Также учитывается, что задача разделения смеси

распределений – это задача максимизации. В целом же структура нового алгоритма для разделения смеси распределений повторяет структуру Алгоритма 6. Сравнительные результаты нового генетического алгоритма для решения серии задач разделения смеси распределений сведены в Таблицу 3.

Как видно, новый алгоритм дает результаты, не уступающие результатам известных алгоритмов, решая одновременно серию задач.

Глава 3 посвящена расширению круга метрик и мер расстояния, которые могут быть задействованы в моделях и алгоритмах автоматической группировки.

Шаг 2 ALA-процедуры предполагает решение задачи Вебера – одной из классических задач теории размещения.

$$X^* = \arg \min_{X \in \mathbb{R}^n} f(X) = \arg \min_{X \in \mathbb{R}^n} \sum_{i=1}^N w_i L(A_i - X).$$

Здесь $A_i \in \mathbb{R}^n, i \in \{1, N\}$ – некоторые заданные точки, называемые точками-потребителями или векторами данных, $w_i \in \mathbb{R}, w_i \geq 0$ – их весовые коэффициенты, $L()$ – некоторая метрика или другая мера расстояния.

Задача Вебера с преградами и запрещенными зонами представляет собой гораздо более сложную задачу оптимизации. Некоторые варианты задачи Вебера представляют собой задачу невыпуклой оптимизации, которую сложно решить точно.

Самыми удобными в этом смысле мерами расстояния являются квадрат евклидова расстояния (не является метрикой) и прямоугольная метрика – в этом случае задача Вебера решается неитеративно алгоритмом линейной сложности.

Если условием решения задачи размещения или автоматической группировки является обязательное пренебрежение небольшими расстояниями, не превышающими некоторого предела (будем считать, что предел равен 1, чего можно достичь нормировкой), имеем меру расстояния $L(X, Y) = \max\{\|X - Y\|, 1\} \forall X, Y$. В работе П.Станимировича и Л.А.Казаковцева показано, что задача с такой мерой расстояния может быть разложена на серию задач размещения с евклидовой метрикой, где область допустимых решений ограничена окружностями и эквивалентна области пересечения окружностей с центрами в точках-потребителях.

Допустимые зоны определены как множество ограничений $\|X - A_i\| - 1 \leq 0, \forall i \in S_<$ и $1 - \|X - A_i\| \leq 0, \forall i \in S_>$, где $S_<, S_> \in \{1, N\}, S_< \cap S_> = \emptyset$, это подмножества множества индексов известных точек (точек-потребителей), а N – общее количество точек-потребителей.

Множество допустимых решений для данной задачи невыпуклое, а целевая функция $f(X)$ выпукла. Решение задачи оптимизации с ограничениями и выпуклой целевой функцией совпадает с решением задачи без ограничений или лежит на границе запрещенной зоны. Таким образом, если X^* является решением задачи, заданной с ограничениями, тогда X^* – решение также и задачи без ограничений, либо $\exists i \in \{1, N\} : \|A_i - X^*\|_2 = 1$. Применение процедуры Вайсфельда, типичной для задачи Вебера без ограничений, может привести к образованию новой точки X^{**} за пределами зоны допустимых решений, заданной ограничениями. Алгоритм, предложенный Л.А.Казаковцевым и П.Станимировичем, основан на замене точки X^* на ближайшую точку в области допустимых решений.

Это единственный известный алгоритм для решения задач оптимизации для такой меры расстояния и в поиске альтернативы мы обратились к одному из перспективных направлений оптимизации – алгоритмам роевого интеллекта, в частности, алгоритму пчелиной колонии.

Алгоритм пчелиной колонии (АПК) имитирует пищевое поведение роя медоносных пчел. Половину популяции виртуальных пчел составляют пчелы-работчие, вторая половина состоит из наблюдателей и разведчиков. Каждый источник пищи – это возможное решение. Для каждого источника пищи есть своя рабочая пчела, которая ищет пищу в окрестности этого источника пищи и передает информацию о качестве источника пищи наблюдателям. Те летят к наиболее качественным источникам и ищут пищу в их окрестности. Работчие с неперспективных источников превращаются в разведчиков и летят на поиски новых источников. В алгоритме пчелиной колонии для задачи оптимизации с ограничениями есть пять управляющих параметров. Наряду с обычными для алгоритмов роевого интеллекта, такими как размер популяции (SP), максимальное количество циклов (MCN) и уровень модификации MR , есть еще два – $limit$ и период разведки (SPP). Они используются в фазе разведчиков.

Для задачи Вебера с ограничениями и допустимыми зонами, ограниченными окружностями, мы задействовали версию АПК с плавающей точкой. Наша реализация алгоритма пытается получить оптимальный вектор $[x_1, x_2]$, сводящий целевую функцию к минимуму и удовлетворяющий ограничениям.

Алгоритм 7. АПК для задачи оптимизации с ограничениями

1. Инициализировать управляющие параметров SP , MCN , MR , $limit$ и SPP . Инициализировать в пространстве поиска популяции решений x_{ij} , $i = 1, \dots, SP/2$, $j = 1, \dots, D$ случайным образом. Вычислить каждое x_i , $i = 1, \dots, SP/2$. Инициализировать параметр $failure_i = 0$, $i = 1, \dots, SP/2$. Присвоить $cycle = 1$.

2. /*Фаза пчел-работчих*/ Для каждого i от 1 до $SP/2$ выполнить

3. Создать новое решение v_i для пчел-работчих x_i . Применить ограничения на созданное решение v_i . Произвести выбор между решениями x_i и v_i .

4. Если решение x_i не улучшено, то присвоить $failure_i = failure_i + 1$, иначе $failure_i = 0$. Следующая итерация 2.

5. /*Фаза пчел-наблюдателей*/ Для каждого i от 1 до $SP/2$

6. Вычислить величины вероятностей P_i для каждого решения x_i :

$P_i = 0.9 * (fitness_i / maxfit) + 0.1$ /* $maxfit$ это лучшее значение целевой функции в популяции */ Следующая итерация 5.

7. Присвоить $t = 1$ и $i = 1$.

8. Пока $t < SP/2$ выполнить

9. Если $S_i < P_i$ то $t = t + 1$ /* S_i случайное вещественное число в диапазоне $[0,1)$ */

10. Создать новое решение v_i для наблюдателей x_i . Применить ограничения на созданное решение v_i . Произвести между решениями x_i и v_i .

11. Если решение x_i не улучшено, то присвоить $failure_i = failure_i + 1$, иначе $failure_i = 0$.

12. Присвоить $i = i + 1$. Если $i = SP/2$, то $i = 1$. Следующая итерация 8.

13. /*Фаза пчел-разведчиков*/ Если $cycle$ кратно SPP , то каждое решение, у которого $failure_i$ больше или равно значению $limit$ заменяется случайно созданным решением.

14. Запомнить лучшее решение на данный момент

15. Присвоить $cycle = cycle + 1$

16. Если $cycle < MCN$, то переход на шаг 2.

Кроме обычного АПК в нашем исследовании мы применили к задаче Вебера с допустимыми зонами, ограниченными окружностями, новый кроссоверный алгоритм пчелиной колонии (КО-АПК), отличающийся уравнением поиска, линейно

возрастающим MR , процедурой исследования окрестности качественного решения и кроссоверным оператором.

В кроссоверном операторе КО-АПК после каждой SPP -ой итерации каждое решение x_i , $failure_i$ которого превышает $limit$, заменяется новым, создаваемым по:

$$v_{ij} = \begin{cases} y_j, & \text{если } R_j < 0.5 \\ x_{ij}, & \text{если } R_j \geq 0.5 \end{cases}$$

где y_j это j -ый элемент лучшего глобального решения, найденного на текущий момент, R_j – случайно выбранное вещественное число из $[0, 1)$, $j = 1, 2, \dots, D$.

Также надо упомянуть, что в КО-АПК используется другой механизм обработки ограничений, отличный от АПК, что позволяет сохранять разнообразие популяции:

$$v_{ij} = \begin{cases} 2 \cdot l_j - v_{ij}, & \text{если } v_{ij} < l_j \\ 2 \cdot u_j - v_{ij}, & \text{если } v_{ij} > u_j \\ v_{ij}, & \text{иначе} \end{cases}$$

где v_{ij} – переменная j решения-кандидата i , l_j и u_j это нижняя и верхняя границы параметра v_{ij} .

Производительность двух метаэвристических алгоритмов и эвристического алгоритма Л.А.Казаковцева и П.С.Станимировича испытывается на 10 тестовых заданиях. Задачи сгенерированы случайным образом, реализованы в вариантах для 5, 10, 50, 100, 250 и 500 точек. Следовательно, эти тестовые задания включают в себя множество ограничений с количеством членов от 5 до 500.

Как видно из Таблицы 4, каждый из метаэвристических алгоритмов выдал лучшие результаты, очень близкие к результатам, полученным с помощью эвристического алгоритма, для каждой из тестовых задач. КО-АПК позволил найти лучшие или равные результаты по сравнению с эвристическим алгоритмом для всех задач, за исключением одной.

Таблица 4. Сравнение наилучших решений АПК и КО-АПК

N	D	АПК	Новый КО-АПК	Эвристич. (Казаковцев- Станимирович)
1	5	38.6309142364	38.6308975913	38.6308975914
2	5	32.8210021279	32.8210021279	32.8210021279
3	10	18.7528484934	18.7528484372	18.9069002076
4	50	723.0047448967	723.0047448967	723.0047301353
5	50	108.9894138815	108.9894138815	114.6249767471
6	50	455.1344639631	455.1344639631	455.1344852622
7	100	231.0855570166	231.0855570165	231.0855570171
8	250	536.8203922596	536.8203916167	536.8203916176
9	500	6193.792794745	6193.792794745	6194.381725440
10	500	4667.0515567517	4667.049269677	4667.482043242

Для корректного сравнения метаэвристических алгоритмов базовые параметры были ограничены до 20 решений и 200 итераций для всех алгоритмов. Для каждого эксперимента произведено 30 запусков. Сходимость эвристического алгоритма подтверждена экспериментально на случайным образом сгенерированных задачах.

Средние значения и значения стандартных отклонений показывают надежность нового алгоритма.

КО-АПК последовательно сходил к одному и тому же решению с тем же значением целевой функции и с очень низким стандартным отклонением. Алгоритм Казаковцева-Станимировича детерминирован, но, как видно из Таблицы 4, иногда дает неточные результаты. Вычислительное время для АПК и нового КО-АПК не слишком отличается друг от друга и для числа ограничений равного 500, разница составляет менее 0.1 секунды.

Таким образом, новый алгоритм КО-АПК позволяет получать наиболее точные решения задачи. Алгоритм может быть использован в составе алгоритма для решения задачи Вебера с мерой расстояния $L(X, Y) = \max\{\|X - Y\|, 1\}$ и в составе алгоритма для решения соответствующей p -медианной задачи (см. таблицы в Главе 2 диссертации).

ЗАКЛЮЧЕНИЕ

В диссертации разработаны новые алгоритмы метода жадных эвристик для одновременного решения серии задач автоматической группировки и размещения объектов с использованием различных метрик и мер расстояния, а также основанных на модели разделения смеси вероятностных распределений, позволяющие эффективно решать широкий круг задач автоматической группировки, к результатам которых предъявляются высокие требования точности и стабильности результата (по значению целевой функции) при большом объеме входных данных и заранее неизвестном числе групп.

Цель диссертации достигнута путем решения поставленных задач, а именно:

1. Проведен анализ существующих методов и алгоритмов решения задач автоматической группировки, размещения и иных задач оптимизации. Выявлено, что методы, предполагающие одновременное с решением задачи автоматической группировки решение задачи определения числа групп в данных, обладают достаточно низкой точностью (по значению целевой функции).

2. Разработаны новые алгоритмы метода жадных эвристик со стратегией глобального поиска, реализованной особым генетическим алгоритмом с гетерогенной (смешанной) популяцией решений, содержащей решения задач с различным числом групп (кластеров) и особой модификацией жадных эвристических процедур. При этом в качестве алгоритмов локального поиска, предусмотренных методом жадных эвристик, могут использоваться различные эффективные алгоритмы локального поиска для задач с соответствующей мерой расстояния. Экспериментально доказано, что разработанные алгоритмы позволяют в случае большой размерности практических задач получать результаты одновременно для серии задач, не уступающие по точности и стабильности результата известным алгоритмам решения единственной задачи, при этом позволяя одновременно решать сразу серию задач с различным числом кластеров. За счет этого повышается эффективность работы систем автоматической группировки без снижения требований к получаемому значению целевой функции.

3. Впервые разработан генетический алгоритм одновременного решения серии задач нечеткой кластеризации данных большой размерности на основе модели разделения смеси вероятностных распределений с различным предполагаемым числом распределений, при заранее известном максимальном числе распределений. Экспериментально доказано, что разработанный алгоритм позволяет в случае большой размерности задач получать результаты одновременно для серии задач, практически не уступающие по достигаемому значению целевой функции и стабильности результата известным алгоритмам решения единственной задачи разделения смеси распределений,

при этом позволяя одновременно решать сразу серию задач с различным числом распределений в смеси. За счет этого повышается эффективность работы систем автоматической группировки без снижения требований к получаемому значению целевой функции правдоподобия.

4. Разработан новый эвристический алгоритм решения задачи Вебера с допустимыми зонами, ограниченными окружностями, который может эффективно применяться в составе алгоритма решения p -медианной задачи с мерой расстояния, ограниченной снизу. Показано, что новый алгоритм позволяет быстро получать значительно более точные результаты в сравнении с единственным известным алгоритмом.

ПУБЛИКАЦИИ ПО ТЕМЕ ДИССЕРТАЦИИ

Публикации в изданиях, рекомендованных ВАК России:

1. Гудыма М.Н. Параллельный алгоритм для p -медианной задачи / Л.А.Казаковцев, А.Н.Антамошкин, М.Н.Гудыма // Системы управления и информационные технологии.–2013.–№ 2.1.–С.124-128.

2. Гудыма М.Н. Алгоритм решения задачи размещения для некоторых специальных метрик / М.Н.Гудыма // Системы управления и информационные технологии.–2013.–№ 4.–С.20-23.

3. Гудыма М.Н. Задача выбора оптимального размещения элементов беспроводной сети / Л.А.Казаковцев, М.Н. Гудыма, А.А. Ступина, Ю.И. Кириллов // Современные проблемы науки и образования [электронное издание].–2013.–вып.3.– URL <http://elibrary.ru/download/29613015.pdf>.

4. Гудыма М.Н. Решение серий задач автоматической группировки промышленной продукции / М.Н. Гудыма, Л.А. Казаковцев, А.Н. Антамошкин // Экономика и менеджмент систем управления.–2016. –Т. 22.–№ 4.–С.80-87.

5. Гудыма М.Н. Стабильность решений серийного алгоритма с жадной эвристикой для задачи автоматической группировки / М.Н.Гудыма, В.И. Орлов, Л.А.Казаковцев, А.Н.Антамошкин, // Системы управления и информационные технологии.–2016. –Т. 66.–№ 4.1.–С.141-145.

6. Гудыма М.Н. Алгоритмы для задачи Вебера с допустимыми зонами, ограниченными окружностями/ М.Н.Гудыма, Л.А.Казаковцев, А.Н.Антамошкин, // Системы управления и информационные технологии.–2017. –Т. 67.–№ 1.–С.4-9.

7. Гудыма М.Н. Эволюционные алгоритмы решения серии задач автоматической группировки с динамическими и гетерогенными популяциями/ М.Н.Гудыма, Л.А.Казаковцев // Системы управления и информационные технологии.–2017. –Т. 68.–№ 2.–С.33-38.

Публикации в изданиях, входящих в систему цитирования Scopus:

8. Gudyma M.N. Algorithms for Location Problems Based on Angular Distances/ L.A. Kazakovtsev et al.// Advances in Operations Research.–2014.– Vol.2014.– Article ID 701267, 12 P.

9. Gudyma M.N. Genetic Algorithm with Greedy Heuristic for Capacity Planning / L.A.Kazakovtsev, M.N.Gudyma, A.N.Antamoshkin // 6th International Congress on Ultra Modern Telecommunications and Control Systems and Workshops (ICUMT) – S.-Petersburg, 2014, 6-8 October .–2014.– P. 607-613.

В других изданиях:

10. Гудыма М.Н. Постановка задачи оптимального размещения сети датчиков мониторинга загрязнения воздуха и воды // Перспективы развития информационных

технологий : тез. докл. междунар. науч.-практ. конф., 24 июня 2013./ Л.А. Казаковцев, М.Н.Гудыма.— Новосибирск.— с. 19-24.

11. Гудыма М.Н. Задача планирования загрузки литейного производства // Актуальные вопросы науки: тез. докл. междунар. науч.-практ. конф., 2014./ М.Н.Гудыма.— М.:— с. 63-64.

12. Гудыма М.Н. Решение задачи Вебера для специальных случаев размещения на плоскости // Решетневские чтения: тез. докл. междунар. науч. конф., 2014./ Л.А. Казаковцев, М.Н.Гудыма.— Красноярск, Сиб. гос. аэрокосмич. ун-т, 2014. — т.2.— №18.— с.52-53.

13. Гудыма М.Н. Использование параллельных вычислений для оптимизации размещения на сети // Проблемы современной аграрной науки: тез. докл. междунар. науч. конф., 2013./ Л.А. Казаковцев, М.Н.Гудыма.— Красноярск, Краснояр. гос. аграр.ун-т.— с.95-97.

14. Гудыма М.Н. Задача нормировки исходных данных испытаний электрорадиоизделий космического применения для алгоритма автоматической группировки / В.В. Федосов, Л.А.Казаковцев, М.Н.Гудыма // Информационные технологии моделирования и управления.—2016. —Т. 100.—№ 4.—С.263-268.

15. Гудыма М.Н. Генетический алгоритм с жадной эвристикой для решения серии задач // Решетневские чтения: тез. докл. междунар. науч. конф., 2016./ М.Н.Гудыма, В.И.Орлов, Л.А. Казаковцев.— Красноярск, Сиб. гос. аэрокосмич. ун-т, 2016. — т.2.— с.28-30.

16. Гудыма М.Н. Применение ЕМ-алгоритма к задаче автоматической группировки электрорадиоизделий // Решетневские чтения: тез. докл. междунар. науч. конф., 2016/ В.И.Орлов, Д.В.Сташков, М.Н.Гудыма, Л.А. Казаковцев.— Красноярск, Сиб. гос. аэрокосмич. ун-т, 2016. — т.2. — с.74-76.

17. Гудыма М.Н. Генетические алгоритмы метода жадных эвристик для серии задач разделения смеси распределений / Д.В.Сташков, М.Н.Гудыма // Информационные технологии моделирования и управления.—2017. —Т. 105.—№ 3.—С.181-191.

Свидетельства о государственной регистрации программ для ЭВМ:

1. Гудыма М.Н. Система поддержки принятия решений об оптимальном размещении элементов беспроводной сети/ Л.А. Казаковцев, М.Н.Гудыма, А.А.Ступина.— М.: РОСПАТЕНТ.—2013.—Свидетельство о государственной регистрации программы для ЭВМ № 2013661595 от 11.12.2013.

2. Гудыма М.Н. Развитие методики ускоренной летной квалификации новых технологий на навигационных космических аппаратах / А.А. Ступина, М.И. Мельдер, Л.А. Казаковцев, М.Н. Гудыма.— М.: РОСПАТЕНТ.— 2013.— Свидетельство о государственной регистрации программы для ЭВМ № 2014610051 от 09.01.2014.

3. Гудыма М.Н. Система поддержки составления графиков загрузки производственных мощностей для непрерывного производства/ Л.А. Казаковцев, М.Н. Гудыма, А.А. Ступина, М.И. Мельдер.— М.: РОСПАТЕНТ.—2013.—Свидетельство о государственной регистрации программы для ЭВМ № 2014610048 от 09.01.2014.