На правах рукописи

Рожнов Иван Павлович

АЛГОРИТМЫ ПОИСКА С ЧЕРЕДУЮЩИМИСЯ РАНДОМИЗИРОВАННЫМИ ОКРЕСТНОСТЯМИ ДЛЯ ЗАДАЧ АВТОМАТИЧЕСКОЙ ГРУППИРОВКИ ОБЪЕКТОВ

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

АВТОРЕФЕРАТ
диссертации на соискание ученой степени
кандидата технических наук

Красноярск – 2019

Работа выполнена в ФГБОУ ВО «Сибирский государственный университет науки и технологий имени академика М.Ф. Решетнева», г. Красноярск.

Научный руководитель:	доктор технических наук, доцент **Казаковцев Лев Александрович**
Официальные оппоненты:	**Крутиков Владимир Николаевич** доктор технических наук, доцент, ФГБОУ ВО «Кемеровский государственный университет», профессор кафедры прикладной математики
	Леванова Татьяна Валентиновна кандидат физ.-мат. наук, доцент, Омский филиал Института математики им. С.Л. Соболева СО РАН, старший научный сотрудник
Ведущая организация:	Федеральное государственное бюджетное образовательное учреждение высшего образования «Воронежский государственный технический университет»

Защита состоится 11 октября 2019 года в 14 часов на заседании диссертационного совета Д 212.249.05, созданного на базе ФГБОУ ВО «Сибирский государственный университет науки и технологий имени академика М.Ф. Решетнева» по адресу: 660037, г. Красноярск, проспект имени газеты Красноярский рабочий, 31.

С диссертацией можно ознакомиться в библиотеке ФГБОУ ВО «Сибирский государственный университет науки и технологий имени академика М.Ф. Решетнева» и на сайте: https://www.sibsau.ru.

Автореферат разослан «____» _____ 2019 г.

Ученый секретарь
диссертационного совета,
канд. техн. наук, доцент Панфилов Илья Александрович

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность. В связи с ускоренным ростом объемов данных растет и потребность в современных средствах и системах сбора, хранения и обработки массивов данных, вследствие чего увеличивается их многообразие. Всё возрастающее использование массивов данных большой размерности стимулирует повышенный интерес к разработке и применению методов и средств обработки и анализа этих массивов. Одним из перспективных направлений является кластерный анализ, который позволяет упорядочивать (группировать) объекты в однородные группы (кластеры), а решение задачи автоматической группировки (кластеризации) сводится к разработке алгоритма, способного обнаружить эти группы без использования предварительно маркированных данных.

Существуют производственные задачи автоматической группировки объектов, которые должны быть решены сравнительно быстро, при этом результат должен быть таким, чтобы известными методами его трудно было бы улучшить без значительного увеличения временных затрат.

Анализ существующих проблем применения методов автоматической группировки объектов, к которым предъявляются высокие требования по точности и стабильности результата, показывает дефицит алгоритмов, способных выдавать за фиксированное время результаты, которые было бы трудно улучшить известными методами, и которые бы обеспечивали стабильность получаемых результатов при многократных запусках алгоритма. При этом известные алгоритмы (например, метода жадных эвристик) требуют значительных вычислительных затрат. Отмечая некоторый дефицит компромиссных по качеству результата и времени счета методов автоматической группировки (под качеством будем понимать точность – близость значения целевой функции к глобальному оптимуму) в настоящем исследовании ставится задача разработать усовершенствованные алгоритмы для задач автоматической группировки, в которых предъявляются высокие требования к точности и стабильности результата.

Степень разработанности темы. В настоящее время в любой дисциплине, предполагающей многомерный анализ данных, существуют задачи автоматической группировки (кластеризации). Существует множество различных методов и алгоритмов автоматической группировки. Наиболее известной моделью кластерного анализа является модель k-средних, которая была предложена Штейнгаузом (1957 г.), а Ллойдом был разработан сам алгоритм. С тех пор алгоритм k-средних, его улучшение, модификация и сочетание с другими алгоритмами, становились темой работ многих исследователей.

В первую очередь среди ученых, в чьих трудах получила развитие автоматическая группировка объектов, необходимо выделить Дюрана Б., Оделла П., Манделя И., Маккуина Дж. Модели автоматической группировки часто имеют сходство с моделями теории размещения объектов, а иногда даже идентичны им, поэтому нередко рассматривались исследователями совместно. Существенный вклад в эти исследования внесли Дрезнер Ц., Хамахер Х., Бримберг Д., Младенович Н. (задачи размещения), Весоловский В. (широкий круг задач), Хакими С. (задачи на сети), Лав Р. (непрерывные задачи с различными метриками). В СССР Хачатуров В.Р. и Черенин В.П. занимались исследованием вопроса размещения предприятий. В Институте математики им. Соболева С.Л. СО РАН работы Гимади Э.Х., Береснева В.Л., Колоколова А.А., а позже Кочетова Ю.А., Еремеева А.В., Забудского Г.Г., Левановой Т.В. и др. при разработке моделей стандартизации и унификации заложили основу для разработки программно-математического аппарата решения задач автоматической группировки и теории размещения объектов.

Метод поиска с чередующимися окрестностями, разработанный Младеновичем Н. и Хансеном П. стал популярным методом решения задач дискретной оптимизации (что отражено в работах Кочетова Ю.А., Лопеса Ф.Г., Бримберна Дж., Левановой Т.В., Алексеевой Е.В. и др.), позволяющим находить хорошие субоптимальные решения достаточно больших задач автоматической группировки.

Казаковцевым Л.А. и Антамошкиным А.Н. предложено применение алгоритмов с жадной эвристической процедурой и показано их преимущество над считающимися классическими алгоритмами автоматической группировки (k-средних, PAM, j-means и др.) для многомерных данных (2014 г.) Сам метод является расширенным подходом к построению процедур псевдобулевой оптимизации и кластеризации. Метод жадных эвристик использует эволюционные алгоритмы как один из возможных способов организации глобального поиска, в том числе подходы и Красноярской школы эволюционных алгоритмов.

Диссертация посвящена разработке и исследованию алгоритмов автоматической группировки с повышенными требованиями к точности и стабильности результата с комбинированным применением алгоритмов поиска с чередующимися рандомизированными окрестностями и жадных эвристических алгоритмов автоматической группировки, в том числе для массивно-параллельных систем.

Основной **идеей** настоящей диссертации является разработка новых алгоритмов метода жадных эвристик с комбинированным применением алгоритмов поиска с чередующимися рандомизированными окрестностями.

Объектом диссертационного исследования являются задачи автоматической группировки многомерных данных, **предметом исследования** – алгоритмы для решения данных задач.

Целью исследования является повышение эффективности систем автоматической группировки объектов, к которым предъявляются высокие требования по точности и стабильности результата (улучшение достигаемого значения целевой функции за заданное время).

Задачи, решаемые в процессе достижения поставленной цели:

1. Анализ существующих проблем при применении методов автоматической группировки объектов, к которым предъявляются высокие требования по точности и стабильности результата.
2. Разработка новых алгоритмов поиска с чередующимися рандомизированными окрестностями и жадных эвристических процедур для задачи k-средних.
3. Разработка новых алгоритмов поиска с чередующимися рандомизированными окрестностями и жадных эвристических процедур для задачи k-медоид.
4. Разработка комбинированного алгоритма на основе классификационного EM-алгоритма (CEM – Classification Expectation Maximization) с применением поиска с чередующимися рандомизированными окрестностями и жадных эвристических процедур.
5. Реализация алгоритмов метода жадных эвристик для задач автоматической группировки для массивно-параллельных систем.
6. Разработка процедуры составления ансамблей алгоритмов автоматической группировки, позволяющей повысить точность разделения (то есть уменьшить число ошибок) сборной партии промышленной продукции на однородные партии промышленной продукции с использованием данных неразрушающих тестовых испытаний.

Методы исследования. Для решения поставленных задач использовались методы системного анализа, исследования операций, теории оптимизации, параллельных вычислений.

Новые научные результаты и положения, выносимые на защиту:

1. Предложен новый подход к разработке алгоритмов автоматической группировки, основанных на параметрических оптимизационных моделях, с комбинированным применением алгоритмов поиска с чередующимися рандомизированными окрестностями и жадных агломеративных эвристических процедур. Показано, что применение данного подхода позволяет создавать эффективные алгоритмы автоматической группировки (по достигаемому значению целевой функции за фиксированное время), основанные на различных оптимизационных моделях.

2. С использованием нового подхода разработаны новые алгоритмы поиска с чередующимися рандомизированными окрестностями для задач k-средних, k-медоид, задачи четкой кластеризации на основе разделения смеси вероятностных распределений (с применением классификационного EM-алгоритма). Продемонстрировано, что новые алгоритмы позволяют получать более точный и стабильный результат (по достигаемому значению целевой функции) в сравнении с известными алгоритмами автоматической группировки за фиксированное время, позволяющее использовать алгоритмы в интерактивном режиме принятия решений для практических задач.

3. Предложены параллельные модификации алгоритмов с жадной агломеративной эвристической процедурой для больших задач автоматической группировки, адаптированные к архитектуре CUDA. Было выявлено, что параллельная реализация алгоритма локального поиска, а также отдельных шагов жадной агломеративной эвристической процедуры позволяет построить алгоритм автоматической группировки с высоким коэффициентом ускорения, сокращающим время расчетов в десятки раз без ухудшения достигаемого значения целевой функции.

4. Предложена процедура составления оптимальных ансамблей алгоритмов автоматической группировки с комбинированным применением генетического алгоритма метода жадных эвристик и согласованной матрицы бинарных разбиений для практических задач. Было выявлено, что точность разделения сборной партии промышленной продукции с особыми требованиями качества на однородные партии, выполненного с применением получаемых ансамблей, выше усредненной точности разделения с применением отдельных алгоритмов, отобранных для составления ансамбля.

Значение для теории. Теоретическая значимость результатов диссертационной работы состоит в разработке нового подхода к созданию алгоритмов автоматической группировки, основанных на параметрических оптимизационных моделях, с комбинированным применением алгоритмов поиска с чередующимися рандомизированными окрестностями и жадных агломеративных эвристических процедур, развивающего метод жадных эвристик, а также процедуры составления оптимальных ансамблей алгоритмов кластеризации.

Практическая ценность нового подхода решения задач автоматической группировки с повышенными требованиями к точности и стабильности результата обусловлена широким диапазоном сфер их применения в задачах кластерного анализа, в том числе непосредственно в практических задачах на производстве, где требуется обеспечение высокой точности разделения производственных партий промышленной продукции на однородные партии по результатам тестовых испытаний.

Практическая реализация результатов. Программная реализация новых алгоритмов и процедуры составления оптимальных ансамблей алгоритмов автоматической группировки была встроена в производственный процесс проведения испытаний электронной компонентной базы космических аппаратов в АО «ИТЦ – НПО ПМ» (г. Железногорск) и в состав «Автоматизированной системы управления технологическим процессом производства анодов», используемой на АО «РУСАЛ

Саяногорск», что позволило обеспечить высокую точность разделения на однородные партии промышленной продукции, сократить время расчетов и требования к вычислительным ресурсам, а также (в АО «ИТЦ – НПО ПМ») обеспечить возможность принятия решений об отборе экземпляров продукции из каждой однородной партии для разрушающего анализа в интерактивном режиме. Основная часть настоящего диссертационного исследования была проведена в рамках государственного задания Министерства образования и науки РФ № 2.5527.2017/БЧ «Методы комбинаторной оптимизации в системах автоматической группировки и классификации».

Апробация работы. Основные положения и результаты диссертационной работы докладывались и обсуждались на международных конференциях и семинарах: «Решетневские чтения» (в 2017 и 2018 годах, г.Железногорск), «Optimization Problems and Their Applications OPTA-2018» (2018 г., г.Омск), «Актуальные проблемы электронного приборостроения АПЭП-2018» (2018 г., г.Новосибирск), «Передовые технологии в аэрокосмической отрасли, машиностроении и автоматизации MIST: Aerospace-2018» (2018 г., г.Красноярск), «Advanced Technologies in Material Science, Mechanical and Automation Engineering» (2019 г., г.Красноярск), «Science and education: experience, problems, development prospects» (2019 г., г.Красноярск) и во всероссийских «Лесной и химический комплексы – проблемы и решения» (2017 г., г.Красноярск), «Системы связи и радионавигации» (2018 г., г.Красноярск). Работа в целом обсуждалась на международном семинаре «Advanced Technologies in Material Science, Mechanical and Automation Engineering» (2019 г., г.Красноярск).

Публикации. Основные теоретические и практические результаты диссертации содержатся в 18 публикациях, среди которых 7 работ в ведущих рецензируемых журналах, рекомендуемых в действующем перечне ВАК, 5 – в международных изданиях, индексируемых в системах цитирования Web of Science и Scopus. Имеется свидетельство о государственной регистрации программы для ЭВМ.

Структура и объем диссертации. Диссертация состоит из введения, 4 глав, заключения и приложений. Она изложена на 176 листах машинописного текста, содержит список литературы из 255 наименований.

ОСНОВНОЕ СОДЕРЖАНИЕ РАБОТЫ

Во **введении** обоснована актуальность, поставлена цель и указаны задачи исследования, научная новизна и практическая значимость работы, изложены методы исследования, сформулированы основные положения, выносимые на защиту.

Глава 1 посвящена анализу текущего состояния и развития методов и задач автоматической группировки во взаимосвязи с задачами теории размещения. Обозначены проблемы, возникающие при решении задач кластеризации объектов с повышенными требованиями к точности и стабильности результата.

Современные методы кластерного анализа предлагают широкий выбор средств выявления разнородных по совокупности параметров групп. Наиболее распространенным из подобных методов является метод k-средних (k-means). Собственно алгоритм k-средних является алгоритмом локальной оптимизации, и его результат зависит от выбора начальных значений (усредненных параметров центров или центроидов групп – кластеров). В то же время, метод выявления различных по параметрам групп объектов должен давать воспроизводимые результаты.

Повысить точность методов автоматической группировки позволяет новый подход к разработке алгоритмов автоматической группировки, основанных на параметрических оптимизационных моделях, с комбинированным применением алгоритмов поиска в чередующихся рандомизированных окрестностях и жадных агломеративных эвристических процедур, изложенный в главах 2 и 3.

Задача k-средних состоит в нахождении таких k центров кластеров $X_1 \ldots X_k$ в d-мерном пространстве, чтобы сумма квадратов расстояний от них до заданных точек A_i ($A_1, \ldots, A_N$) была минимальна:

$$\arg\min F(X_1,\ldots,X_k) = \sum_{i=1}^{N} \min_{j\in\{1,k\}} \left\| X_j - A_j \right\|^2, \qquad (1)$$

Одноименный алгоритм последовательно улучшает известное решение, позволяя найти локальный минимум. Это простой и быстрый алгоритм, применимый к широчайшему классу задач. Алгоритм имеет ограничения – в начале решения необходимо задать число групп k, на которые разбиваются объекты, а результат сильно зависит от начального решения, как правило, выбираемого случайно.

Алгоритм 1.1 Алгоритм k-средних

Дано: векторы данных $A_1 \ldots A_N$, k начальных центров кластеров $X_1 \ldots X_k$

выполнять

1: Составить кластер C_i векторов данных для каждого центра X_i так, чтобы для каждого вектора данных его центр был ближайшим.

2: Рассчитать новое значение центра X_i для каждого кластера.

пока шаги 1-2 приводят к каким-либо изменениям.

Алгоритм для задачи k-медоид – Partitioning Around Medoids (PAM) – был предложен Кауфманом (Leonard Kaufman) и Руссивом (Peter J. Rousseeuw) [106]. Он похож на алгоритм k-средних: работа обоих основана на попытках минимизировать ошибку, но PAM работает с медоидами – объектами, являющимися частью исходного множества и представляющими группу, в которую они включены, а k-means работает с центроидами – искусственно созданными объектами, представляющими кластер. Алгоритм PAM разделяет множество из N объектов на k кластеров (N и k являются входными данными алгоритма). Алгоритм работает с матрицей расстояний, его цель – минимизировать расстояние между представителями каждого кластера и его членами.

Алгоритм 1.2 PAM-процедура

Фаза Build:

1. Выбрать k объектов в качестве медоид.

2. Построить матрицу расстояний, если она не была задана.

3. Отнести каждый объект к ближайшей медоиде.

Фаза Swap:

4. Для каждого кластера найти объекты, снижающие суммарное расстояние, и если такие объекты есть, выбрать те, которые снижают его сильней всего, в качестве медоид.

5. Если хотя бы одна медоида поменялась, вернуться к шагу 3, иначе завершить алгоритм.

В число популярных методов автоматической группировки входит и алгоритм Expectation Maximization (EM-алгоритм – максимизация математического ожидания). Основная идея алгоритма состоит в искусственном введении вспомогательного вектора скрытых переменных, что сводит сложную оптимизационную задачу к двум шагам:

1. E-шаг – последовательность итераций по пересчёту скрытых переменных по текущему приближению вектора параметров;

2. M-шаг – максимизации правдоподобия (для нахождения следующего приближения вектора).

Классификационный EM-алгоритм (Classification Expectation Maximization – CEM) – модификация EM-алгоритма работает по принципу четкой классификатора данных выборки. В этом случае каждый объект относится к единственному кластеру. CEM-алгоритм почти совпадает с другой модификацией – SEM (Stochastic EM), только у

первого на каждом шаге вводится детерминированное правило, что данные относятся лишь к одному кластеру, для которого вычислили максимальную апостериорную вероятность. Таким образом, CEM-алгоритм, в отличие от EM, решает задачу четкой кластеризации.

Для дальнейшего изложения в настоящей диссертации каждую из процедур: k-средних, k-медоид и CEM обозначим как двухшаговый алгоритм локального поиска. Тем более что сами процедуры k-средних и k-медоид являются, по сути, алгоритмами поиска с чередующимися окрестностями. Далее при решении задач k-средних в качестве двухшагового алгоритма локального поиска будет реализован Алгоритмом 1.1, соответственно для k-медоид – Алгоритмом 1.2 и максимизации функции правдоподобия – CEM-алгоритмом.

Методы локального поиска получили дальнейшее развитие в метаэвристиках (методах оптимизации, многократно использующих простые правила или эвристики для достижения оптимального или субоптимального решения, характеризующихся большей устойчивостью). Поиск с чередующимися окрестностями (VNS – Variable Neighborhoods Search) – популярный метод решения задач дискретной оптимизации Н. Младеновича и П. Хансена, который позволяет находить хорошие субоптимальные решения достаточно больших задач автоматической группировки. Идея состоит в систематическом изменении вида окрестности в ходе локального поиска. Гибкость и высокая эффективность объясняют ее конкурентоспособность при решении NP-трудных задач.

Обозначим через N_k, k=1,..k_{max}, конечное множество видов окрестностей, предварительно выбранных для локального поиска. Метод с чередующимися окрестностями опирается на то, что локальный минимум в одной окрестности не обязательно является локальным минимумом в другой окрестности, при этом глобальный минимум является локальным в любой окрестности. Кроме того, в среднем локальные минимумы ближе к глобальному, чем случайно выбранная точка, и они расположены близко друг к другу. Это позволяет сузить область поиска глобального оптимума, используя информацию об уже обнаруженных локальных оптимумах. Эта гипотеза лежит в основе различных операторов скрещивания (crossover) для генетических алгоритмов и других подходов.

Реализация метода локального поиска с чередующимися окрестностями возможна одним из трех способов: детерминированным, вероятностным или смешанным, сочетающим в себе два предыдущих. В детерминированном локальном спуске с чередующимися окрестностями (VND) предполагается фиксированный порядок смены окрестностей и поиск локального минимума относительно каждой из них. Вероятностный локальный спуск с чередующимися окрестностями (RVNS) отличается от предыдущего метода VND случайным выбором точек из окрестности $O_k(x)$. Этап поиска лучшей точки в окрестности опускается. Алгоритмы RVNS наиболее продуктивны при решении задач большой размерности, когда применение детерминированного варианта требует слишком много «машинного» времени для выполнения одной итерации.

Основная схема локального поиска с чередующимися окрестностями (VNS) является комбинацией двух предыдущих вариантов (VND и RVNS).

Алгоритм 2 VNS **(**Variable Neighborhoods Search)

Шаг 1. Выбрать окрестности O_k, k=1,..k_{max}, и начальную точку x.

Шаг 2. Повторять, пока не выполнен критерий остановки.

2.1. k=1.

2.2. Повторять до тех пор, пока $k \le k_{max}$:

2.2.1. случайно выбрать точку $x' \in O_k(x)$;

2.2.2. применить локальный спуск с начальной точки x', не меняя координат, по

которым x и x' совпадают. Полученный локальный оптимум обозначается x'';

2.2.3. если $F(x'') < F(x)$, то полагается $x=x''$, $k=1$, иначе $k= k + 1$.

Критерием остановки может служить максимальное время счета или максимальное число итераций без смены лучшего найденного решения. При решении задач большой размерности сложность выполнения одной итерации становится весьма большой, и требуются новые подходы для разработки эффективных методов локального поиска.

Ниже мы рассмотрим VNS-подобные алгоритмы автоматической группировки, основанные на параметрических оптимизационных моделях, с применением жадных агломеративных эвристических процедур.

Жадная агломеративная эвристическая процедура для задачи k-средних и аналогичных задач состоит из двух шагов. Пусть имеются два известных (родительских) решения задачи (первое из которых, например, является лучшим из известных), представленных множествами центров кластеров S. Вначале множества родительских решений объединяются. Получаем промежуточное недопустимое (с избыточным числом кластеров) решение. Затем производится последовательное уменьшение числа центров. Каждый раз отсекается тот центр, удаление которого даёт наименее существенное ухудшение значения целевой функции. Ниже представлен алгоритм работы базовой жадной агломеративной эвристической процедуры:

Алгоритм 3.1 Базовая жадная агломеративная эвристическая процедура (Greedy)

Дано: начальное число кластеров K, требуемое число кластеров $k<K$.

1: Случайным образом определить начальное решение $S=\{X_1,\ldots,X_k\}$.

2: Передать решение S в двухшаговый алгоритм локального поиска, получить новое, улучшенное решение S.

пока $K\neq k$

для каждого $i' \in \overline{\{1,K\}}$

3.1: $S' = S \setminus \{X_{i'}\}$.

3.2: Передать решение S' в двухшаговый алгоритм локального поиска, выполнить от 1 до 3 итераций алгоритма, полученное значение (значение целевой функции) сохранить в $F'_{i'}$.

конец цикла

5: $i'' = \arg\max\limits_{i'=1,k} F_{i'}$.

6: Получить решение $S'' = S \setminus \{X_{i''}\}$, улучшить его с помощью двухшагового алгоритма локального поиска.

конец цикла

Глава 2 посвящена разработке комбинированных алгоритмов метода жадных эвристик для задач автоматической группировки с повышенными требованиями к точности и стабильности результата, с совместным применением алгоритмов поиска с чередующимися рандомизированными окрестностями, а также параллельных жадных эвристических алгоритмов автоматической группировки для массивно-параллельных систем применительно к задаче k-средних.

Производительность Алгоритма 3.1 при больших объемах данных во время выполнения расчетов становится проблемой, особенно когда найти правильный параметр k (число кластеров) в практических задачах можно только путем выполнения нескольких запусков с разным количеством кластеров. При увеличении количества кластеров Алгоритм 3.1 на шаге 2 начинает работать всё медленнее (алгоритм требует все большего количества итераций, и каждая итерация требует возрастающих вычислительных

ресурсов), поэтому мы внесли изменения и реализовали удаление кластеров не по одному, а по несколько за одну итерацию.

Алгоритм 3.2 Базовая жадная агломеративная эвристическая процедура для задач с большим числом кластеров

Дано: начальное число кластеров K, необходимое количество кластеров $k<K$, $k>50$, первоначальное решение S, $|S|=K$.

1: Улучшить решение S двухшаговым алгоритмом локального поиска (если это возможно).

пока $K \neq k$

для каждого $i' \in \{\overline{1,K}\}$

2: $S' = S \setminus \{X_{i'}\}$. Вычислить $F'_{i'} = F(S')$, где $F(.)$ значение целевой функции (например, (1) для задачи k-средних).

конец цикла

3. Установить S_{elim} из n_{elim} центроидов, $S_{elim} \subset S$, $|S_{elim}|=n_{elim}$, с минимальными значениями $F'_{i'}$. Здесь, $n_{elim}=\max\{1, 0.2\cdot(|S|-k)\}$.

4: Получить новое решение $S=S\setminus S_{elim}$, $K=K-1$, и улучшить его с помощью двухшагового алгоритма локального поиска.

конец цикла

Предложены новые эвристические процедуры, модифицирующие известное решение на основе второго известного решения (см. Алгоритмы 3.1 и 3.2).

Алгоритм 4 Жадная процедура 1

Дано: множества центров кластеров $S'=\{X'_1,...,X'_k\}$ и $S''=\{X''_1,...,X''_k\}$

для каждого $i' \in \{\overline{1,K}\}$

1: Объединить S' с элементом множества S'': $S = S' \cup \{X''_{i'}\}$

2: Запустить базовую жадную агломеративную эвристическую процедуру (Алгоритм 3.1 или 3.2) с S в качестве начального решения. Полученный результат (полученное множество, а также значение целевой функции) сохранить.

3: Возвратить в качестве результата лучшее (по значению целевой функции) из решений, полученных на шаге 2.

конец цикла

Возможен вариант, в котором множества объединяются частично, при этом первое множество берется полностью, а из второго множества выбирается случайным образом случайное число элементов.

Алгоритм 5 Жадная процедура 2

Дано: см. Алгоритм 4.

1: Выбрать случайное $r' \in [0;1)$. Присвоить $r=[(k/2-2)\, r'^2]+2$.

Здесь [.] – целая часть числа.

2: **для** i **от** 1 **до** $k-r$

2.1: Сформировать случайно выбранное подмножество S''' элементов множества S'' мощности r. Объединить множества $S = S' \cup S'''$.

2.2: Запустить базовую жадную агломеративную эвристическую процедуру (Алгоритм 3.1 или 3.2) с этими объединёнными множествами в качестве начального решения.

конец цикла

3: Возвратить в качестве результата лучшее (по значению целевой функции)

из решений, полученных на шаге 2.2.

Более простой вариант Алгоритма 4 представлен ниже, уже с полным объединением множеств.

Алгоритм 6 Жадная процедура 3

Дано см. Алгоритм 4.

1: Объединить множества $S = S' \cup S''$.

2: Запустить базовую жадную агломеративную эвристическую процедуру (Алгоритм 3.1 или 3.2) с *S* в качестве начального решения.

Данные алгоритмы могут использоваться в составе различных стратегий глобального поиска, а в качестве окрестностей, в которых производится поиск решения, используются множества решений, производные («дочерние») по отношению к решению *S'*, образованные комбинированием его элементов с элементами некоторого решения *S''* и применением базовой жадной агломеративной эвристической процедуры (Алгоритм 3.1 или 3.2).

Алгоритмы 4-6 могут осуществлять поиск в окрестностях известного промежуточное решение *S'*, где решения, принадлежащие окрестности, образуются добавлением элементов другого решения *S''* с последующим удалением «лишних» центров кластеров жадной агломеративной эвристической процедурой. Таким образом, второе решение *S''* является параметром окрестности, выбираемым случайным образом (рандомизированным).

Алгоритм автоматической группировки для задачи k-средних с комбинированным применением алгоритмов поиска с чередующимися рандомизированными окрестностями и жадных агломеративных эвристических процедур можно описать следующим образом:

Алгоритм 7.1 k-GH-VNS (Greedy Heuristic in the Variable Neighborhood Search) для задачи k-средних

1: Получить решение *S*, запустив Алгоритм 1.1 из случайным образом сгенерированного начального решения.

2: $O=O_{start}$ номер окрестности поиска.

3: $i=0$, $j=0$.

пока $j < j_{max}$

пока $i < i_{max}$

4: **если** не выполняются условия ОСТАНОВа, **то** получить решение *S'*, запустив Алгоритм 1.1 из случайного начального решения.

повторять

5: В зависимости от значения *S* (возможны значения 1, 2 или 3), запустить Алгоритм Жадная процедура 1, 2 или 3 соответственно с начальными решениями *S* и *S'*. Так, окрестность определяется способом включения центров кластеров из второго известного решения и параметром окрестности – вторым известным решением.

если новое решение лучше, чем *S*, **то**

записать новый результат в *S*, $i=0$, $j=0$.

иначе выйти из цикла.

конец цикла

6: $i=i+1$.

конец цикла

7: $i=0$, $j=j+1$, $O=O+1$, **если** $O>3$, **то** $O=1$.

конец цикла

В данном алгоритме i_{max} – число безрезультатных поисков в окрестности, а j_{max} – число безрезультатных переключений окрестностей. Значения этих двух параметров важны при расчётах. Мы использовали значения: $i_{max}=2k$, $j_{max}=2$.

Параметр O_{start}, задающий номер окрестности, с которой начинается поиск, особенно важен. В зависимости от значения O_{start} далее алгоритмы обозначены GH-VNS1, GH-VNS2, GH-VNS3 (для задачи k-средних соответственно – k-GH-VNS1, k-GH-VNS2, k-GH-VNS3).

Важным является способ получения второго решения *S*’ на Шаге 4. По умолчанию второе решение содержит число центров, равное числу центров в решении *S*. Мы также использовали модификации Алгоритма 7.1, в которых число центров в решении *S*’ выбирается случайным образом из множества $\{\overline{2,|S|}\}$, где |*S*| – число центров в решении *S*. В этом случае алгоритмы названы GH-VNS1-RND, GH-VNS2-RND, GH-VNS3-RND.

Для нашего исследования мы использовали классические наборы данных из репозиториев UCI (Machine Learning Repository) и Clustering basic benchmark. Для всех наборов данных было выполнено по 30 попыток запуска каждого из алгоритмов. Фиксировались только лучшие результаты, достигнутые в каждой попытке, затем из этих результатов по каждому алгоритму были рассчитаны значения целевой функции: минимальное и максимальное значения (Min, Max), среднее значение (Среднее) и среднеквадратичное отклонение (СКО). Алгоритмы j-means и k-средних были запущены в режиме мультистарта.

Лучшие значения целевой функции (минимальное значение, среднее значение и среднеквадратичное отклонение) выделены полужирным курсивом (Таблицы 1-2).

Таблица 1 – Результаты вычислительных экспериментов по набору данных BIRCH3 (100000 векторов данных, каждый размерностью 2) 100 кластеров, 6 часов

Алгоритм	Значение целевой функции			
	Min	Max	Среднее	СКО
k-средних	7,92474E+13	8,87404E+13	8,31599E+13	3,088140E+12
j-means	3,76222E+13	3,7965E+13	3,77715E+13	0,116211E+12
k-GH-VNS1	***3,72537E+13***	3,77474E+13	3,74703E+13	0,171124E+12
k-GH-VNS2	4,21378E+13	5,01871E+13	4,52349E+13	4,333462E+12
k-GH-VNS3	***3,72525E+13***	3,74572E+13	***3,73745E+13***	***0,074315E+12***
k-GH-VNS1-RND	***3,72541E+13***	3,77687E+13	3,74943E+13	0,185483E+12
k-GH-VNS2-RND	3,83257E+13	4,61847E+13	4,0815E+13	2,543163E+12
k-GH-VNS3-RND	3,73131E+13	3,75242E+13	3,74164E+13	***0,061831E+12***

Таблица 2 – Результаты вычислительных экспериментов по набору данных KDDCUP04BioNormed (145751 векторов данных, каждый размерностью 74) 200 кластеров, 24 часа

Алгоритм	Значение целевой функции			
	Min	Max	Среднее	СКО
k-средних	5 336 446	5 381 386	5 366 144	25 722,4
j-means	5 330 344	5 382 908	5 355 903	26 785,8
k-GH-VNS1	***5 294 620***	5 307 828	***5 301 224***	***9 339,5***
k-GH-VNS2	5 440 814	5 476 140	5 458 477	24 979,5
k-GH-VNS1-RND	***5 310 067***	5 340 849	5 325 458	21 765,7
k-GH-VNS2-RND	5 368 527	5 399 695	5 384 111	22 039,6

Для более полного сравнения полученных результатов вычислительных экспериментов новых алгоритмов были использованы данные вычислительных

экспериментов полученных ранее Казаковцевым Л.А. над наборами данных тестовых испытаний партии изделий промышленной продукции различными модификациями генетического алгоритма (Таблица 3). Для расчетов были использованы сборные партии изделий промышленной продукции 1526TL1 – 3 партии (1234 векторов данных, каждый размерностью 157).

В Таблице 3 были использованы следующие аббревиатуры и сокращения: ГА – генетический алгоритм, ЖЭ – жадная эвристика, ГАЖЭ – генетический с жадной эвристикой с вещественным алфавитом, ЛП – локальный поиск, ГА ФП – генетический алгоритм с рекомбинацией подножеств фиксированной длины.

Таблица 3 – Результаты вычислительных экспериментов над результатами испытаний производственных партий изделий 1526TL1 (10 кластеров, 1 минута)

Алгоритм	Значение целевой функции			
	Min	Max	Среднее	СКО
k-средних	43 842,10	43 844,66	43 843,38	0,8346
j-means	***43 841,97***	43 843,51	43 842,59	0,4487
k-GH-VNS1	***43 841,97***	43 844,18	***43 842,34***	0,9000
k-GH-VNS2	***43 841,97***	43 844,18	43 843,46	1,0817
k-GH-VNS3	***43 841,97***	43 842,10	***43 841,99***	***0,0424***
ГАЖЭ+ЛП	43 842,10	43 845,73	43 843,72	1,3199
ГАЖЭ вещ., σ e =0.25	***43 841,98***	43 844,18	43 842,6	0,6762
ГАЖЭ вещ.частич., σ e=0,25	***43 841,98***	43 841,98	***43 841,98***	***1,53E-11***
ГА ФП	***43 841,98***	43 842,34	43 842,10	***0,0945***
ГА классич.	43 842,10	43 842,88	43 842,44	0,2349
Детерм. ЖЭ, σ e =0.25	45 113,56	45 113,56	45 113,56	0,0000
Детерм. ЖЭ, σ e =0.001	45 021,21	45 021,21	45 021,21	0,0000

Ниже представлена реализация алгоритмов метода жадных эвристик с применением архитектуры CUDA и исследование их свойств при решении задач большой размерности. CUDA (от английского Compute Unified Device Architecture) – платформа параллельных вычислений и модель программирования, специально разработанная фирмой NVIDIA для общих вычислений на графических процессорах (GPU), которая позволяет существенно увеличить вычислительную производительность. Графический процессор рассматривается как набор мультипроцессоров, выполняющих параллельные потоки параллельно. Потоки сгруппированы в блоки данных и выполняют те же инструкции по разным данным параллельно. Один или несколько блоков напрямую связаны с аппаратным мультипроцессором, где распределение времени определяет порядок выполнения. Внутри одного блока потоки могут быть синхронизированы в любой точке выполнения.

Мы использовали следующий вариант реализации Шага 1 Алгоритма 1.1. Для первой части параллельного алгоритма, которая реализует 1-й шаг Алгоритма 1.1, мы использовали один поток вычислений (фактически без распараллеливания).

Алгоритм 1.1а CUDA реализация шага 1 Алгоритма 1.1, часть 1

X'_j=0 для всех $j \in \{\overline{1,k}\}$. // Здесь, X'_j векторы, используемые для расчета новых кластерных центров.

$counter_j$=0 для всех $j \in \{\overline{1,k}\}$. // счетчики объектов для каждого кластера

Для второй части алгоритма, которая реализует 1-й шаг Алгоритма 1.1, мы использовали $N_{trreads}$ = 512 потоков для каждого блока CUDA. Количество блоков рассчитывается как:

$N_{blocks}=(N+N_{threads}-1)/N_{threads}$. (2)

Таким образом, каждый поток обрабатывает только один вектор данных.

Алгоритм 1.1b CUDA реализация шага 1 Алгоритма 1.1, часть 2

i = blockIdx.x × blockDim.x + threadIdx.x .
Если $i>N$ тогда возврат.
$j'=\arg\min_j \left\|A_j - X_i\right\|^2$. // номер кластера
$X'_{j'}=X'_{j'}+A_i$.
$C_i=j'$. // приписать A_i для кластера j'.
$counter_{j'}=counter_{j'}+1$.
Синхронизировать потоки.

Для части алгоритма, которая реализует 2-й шаг Алгоритма 1.1, мы использовали $N_{trreads}$ = 512 потоков для каждого блока CUDA. Количество блоков рассчитывается как $N_{blocks2} = (k + N_{threads}-1) / N_{threads}$.

Алгоритм 1.1c CUDA реализация шага 2 Алгоритма 1.1

*j = blockIdx.x * blockDim.x + threadIdx.x* .
Если $j>k$ тогда возврат.
$X_{j'}= X'_i/counter_j$.
Синхронизировать потоки.

Алгоритм 3.2 предполагает многократный запуск алгоритма k-средних (или другого метода локального поиска), и число этих запусков растет с ростом числа кластеров (квадратичная зависимость). Мы предлагаем использовать оптимизированную для GPU стратегию для k-средних, а также адаптированную к архитектуре CUDA процедуру исключения кластеров из решения, которая является обязательным и наиболее вычислительно затратным шагом в жадной агломеративной эвристической процедуре. Для этого мы реализовали Шаг 2 Алгоритма 3.2 на графическом процессоре (GPU). На этом этапе Алгоритм 3.2 вычисляет общее расстояние после удаления одного кластера: $F'_{i'} = F(S')$, где $S' = S \setminus \{X_{i'}\}$. Вычислив $F(S)$, мы можем рассчитать $F'_{i'} = F(S') = F(S) + \sum_{l=1}^{N} \Delta D_l \cdot$, где

$$\Delta D_l . = \begin{cases} 0, & C_{i'} \neq l, \\ \left(\min_{j\in\{\overline{1,k}\},\, j\neq i'} \left\|A_j - X_j\right\|^2\right) - \left\|A_j - X_{C_{i'}}\right\|^2, & C_{i'} = l. \end{cases} \quad (3)$$

где l – номер кластера. Здесь мы использовали 512 потоков для каждого блока CUDA, количество блоков рассчитывается в соответствии с (2). Сначала переменная *sumD* инициализируется со значением 0. Затем для каждого вектора данных запускается следующий алгоритм и вычисляется ΔD_l .

Алгоритм 3.2a CUDA реализация шага 2 Алгоритма 3.2

l = blockIdx.x × blockDim.x + threadIdx.x .
Если $l>k$ тогда возврат.
Рассчитать ΔD_l в соответствии с (3).
Если $\Delta D_l >0$ то atomicAdd(*sumD*, ΔD_l).
Синхронизировать потоки.

Все остальные алгоритмы работают на центральном процессоре (CPU).

Вычислительные эксперименты проводились на тестовой системе: Intel Core 2 Duo E8400CPU, 4GBRAM. Графический процессор NVIDIA GeForce 9600 GT, с 2048 МБ ОЗУ. Лучшие значения целевой функции (минимальное значение, среднее значение и СКО) выделены полужирным курсивом (Таблица 4).

Таблица 4 – Сравнение результатов работы алгоритмов на CPU и GPU по набору данных BIRCH3 (100 кластеров, 30 попыток)

Алгоритм	Значение целевой функции			
	Min	Max	Среднее	СКО
GPU 1 минута				
k-средних	8,18676E+13	9,96542E+13	8,98255E+13	8,37212E+12
j-means	5,30805E+13	13,2286E+13	7,91183E+13	28,2000E+12
k-GH-VNS1	***3,71973E+13***	3,76732E+13	3,73639E+13	0,18509E+12
k-GH-VNS2	3,73240E+13	4,06161E+13	3,91485E+13	1,14305E+12
k-GH-VNS3	3,72082E+13	3,72550E+13	***3,72422E+13***	***0,01998E+12***
GPU 10 минут				
k-средних	7,98405E+13	9,96542E+13	8,93187E+13	9,04845E+12
j-means	4,03266E+13	4,5392E+13	4,23065E+13	1,77787E+12
k-GH-VNS1	***3,71474E+13***	3,71933E+13	***3,71778E+13***	***0,02348E+12***
k-GH-VNS2	***3,71474E+13***	3,72261E+13	3,71834E+13	0,02595E+12
k-GH-VNS3	***3,71473E+13***	3,72453E+13	3,71817E+13	0,03723E+12
CPU 6 часов				
k-средних	7,92474E+13	8,87404E+13	8,31599E+13	3,088140E+12
j-means	3,76222E+13	3,7965E+13	3,77715E+13	0,116211E+12
k-GH-VNS1	3,72537E+13	3,77474E+13	3,74703E+13	0,171124E+12
k-GH-VNS2	4,21378E+13	5,01871E+13	4,52349E+13	4,333462E+12
k-GH-VNS3	***3,72525E+13***	3,74572E+13	***3,73745E+13***	***0,074315E+12***

В исследованиях по набору данных BIRCH3 без использования технологии CUDA (Таблица 1) было получено лучшее минимальное значение целевой функции 3.72525E+13 при условии 6 часов на каждую попытку. При расчётах с использованием графического процессора мы получили лучшее значение целевой функции за 1 минуту. Времени затрачено в 360 раз меньше без снижения точности.

Алгоритмы кластеризации, которые показывают лучшие результаты целевой функции с небольшим числом кластеров, не всегда являются лучшими с увеличением числа кластеров. Однако преимущество семейства жадных эвристических алгоритмов над алгоритмом k-средних, а так же j-means (считающимся одним из лучших) остается после перехода к архитектуре CUDA. Использование графического процессора показывает преимущество в достижении скорости по сравнению с вычислениями на процессоре, и преимущество увеличивается для больших наборов данных и большого количества кластеров в десятки и сотни раз.

Глава 3 посвящена разработке комбинированных алгоритмов метода жадных эвристик для задач автоматической группировки с повышенными требованиями к точности и стабильности результата с применением алгоритмов поиска с чередующимися рандомизированными окрестностями применительно к более широкому кругу задач: задаче k-медоид и максимизации функции правдоподобия математического ожидания.

Применительно к задаче k-медоид и максимизации математического ожидания были разработаны алгоритмы, подобные Алгоритму 7.1, в которых Алгоритм 1.1 заменяется соответственно Алгоритмом 1.2 и CEM-алгоритмом. Результаты вычислительных экспериментов представлены в Таблицах 5 и 6. Лучшие значения целевой функции (минимальное значение, среднее значение и среднеквадратичное отклонение) выделены полужирным курсивом.

Таблица 5 – Результаты вычислительных экспериментов по набору данных ionosphere (351 векторов данных, каждый размерностью 35) 10 кластеров, 60 секунд, 30 попыток, манхэттенское расстояние

Алгоритм	Значение целевой функции		
	Min (рекорд)	Среднее	СКО
PAM	2 688,57	2 704,17	12,3308
PAM-GH-VNS1	***2 607,21***	***2 607,25***	***0,1497***
PAM-GH-VNS2	***2 607,21***	***2 607,43***	***0,4303***
PAM-GH-VNS3	***2 607,21***	***2 607,34***	***0,4159***
GA-FULL	***2 608,22***	2 624,97	9,5896
GA-ONE	***2 608,69***	2 625,18	10,7757

В Таблице 5 GA-FULL – генетический алгоритм с жадной эвристикой с вещественным алфавитом, GA-ONE – генетический алгоритм, в котором Алгоритм 5 (Жадная процедура 2) используется в качестве процедуры кроссинговера.

Таблица 6 – Результаты вычислительных экспериментов по наборам данных тестовых испытаний партии изделий промышленной продукции (10 кластеров, 2 минуты, 30 попыток)

Алгоритм	Значение целевой функции			
	Min	Max	Среднее	СКО
3OT122A (767 векторов данных, каждый размерностью 13)				
CEM	120 947,6	146 428,5	135 777,6	***7 985,6992***
CEM-GH-VNS1	121 256,5	152 729,1	***143 956,0***	8 708,6293
CEM-GH-VNS2	123 664,4	158 759,2	***143 028,5***	10 294,3992
CEM-GH-VNS3	***128 282,2***	155 761,9	***143 506,9***	10 058,8266
1526TL1 (1234 векторов данных, каждый размерностью 157)				
CEM	354 007,3	416 538,4	384 883,4	***20 792,8068***
CEM-GH-VNS1	376 137,1	477 124,5	438 109,4	29 964,0641
CEM-GH-VNS2	345 072,6	487 498,3	444 378,1	43 575,3282
CEM-GH-VNS3	***379 352,3***	516 777,8	***456 271,4***	38 323,0246

Общую схему предлагаемого нового подхода к разработке алгоритмов автоматической группировки, основанных на параметрических оптимизационных моделях, с комбинированным применением алгоритмов поиска с чередующимися рандомизированными окрестностями и жадных агломеративных эвристических процедур можно описать следующим образом:

Алгоритм 7.2 GH-VNS (Greedy Heuristic in the Variable Neighborhood Search)

1: Получить решение *S*, запустив двухшаговый алгоритм локального поиска из случайным образом сгенерированного начального решения.

2: $O=O_{start}$ **(**номер окрестности поиска).

3: $i=0$, $j=0$ (количество безрезультатных итераций в конкретной окрестности и в целом по алгоритму).

пока $j < j_{\max}$

пока $i < i_{\max}$

4: **если** не выполняются условия ОСТАНОВа (превышение лимита времени), **то** получить решение *S'*, запустив двухшаговый алгоритм локального поиска из случайного начального решения.

повторять

5: В зависимости от значения S (возможны значения 1, 2 или 3), запустить Алгоритм Жадная процедура 1 или 2 или 3 соответственно с начальными решениями S и S'. Так, окрестность определяется способом включения центров кластеров из второго известного решения и параметром окрестности – вторым известным решением.

если новое решение лучше, чем S, **то**

записать новый результат в S, i=0, j=0.

иначе выйти из цикла.

конец цикла

6: i=i+1.

конец цикла

7: i=0, j=j+1, O=O+1, **если** O>3, **то** O=1.

конец цикла

Результаты вычислительных экспериментов показали, что новые алгоритмы метода жадных эвристик для задач автоматической группировки с повышенными требованиями к точности результата (по значению целевой функции), с применением алгоритмов поиска с чередующимися рандомизированными окрестностями (GH-VNS) имеют более стабильные (меньшее среднеквадратичное отклонение целевой функции) и более точные (меньшее среднее значение целевой функции) результаты, и следовательно, лучшие показатели в сравнении с классическими алгоритмами (k-средних, j-means, PAM и CEM).

В то же время, с ростом числа кластеров и объема выборки сравнительная эффективность нового подхода, основанного на параметрических оптимизационных моделях, с комбинированным применением алгоритмов поиска с чередующимися рандомизированными окрестностями и жадных агломеративных эвристических процедур, повышается, и для больших наборов данных новые алгоритмы имеют преимущество при фиксированном времени работы алгоритма.

Однако стоит отметить, что при значительном увеличении времени расчетов известные генетические алгоритмы метода жадных эвристик показывают немного лучше результаты в сравнении с предложенными новыми алгоритмами. Тем не менее, можно говорить о конкурентоспособности новых алгоритмов как в сравнении с классическими алгоритмами k-средних, PAM и j-means, так и с генетическими алгоритмами, включая алгоритмы метода жадных эвристик, а также с детерминированными алгоритмами.

Глава 4 посвящена описанию задачи выделения однородных партий для формирования электронной компонентной базы космического применения (как примеру актуальной задачи автоматической группировки с повышенными требованиями к точности и стабильности результата) и разработке процедуры составления оптимальных ансамблей алгоритмов автоматической группировки с комбинированным применением генетического алгоритма метода жадных эвристик и согласованной матрицы бинарных разбиений, позволяющей повысить точность разделения на однородные партии продукции для практических задач автоматической группировки промышленной продукции с применением изложенного в главах 2 и 3 подхода к разработке алгоритмов автоматической группировки.

В работах Орлова В.И., Федосова В.В., Казаковцева Л.А. и др. показано, что поставляемые партии промышленной продукции (на примере интегральных схем) могут быть неоднородными по составу. Для того, чтобы распространить результаты выборочных разрушающих испытаний на всю производственную партию изделий, необходимо быть уверенными в том, что мы имеем дело с партией изделий, изготовленной из единой (однородной) партии сырья или, что разброс параметров невелик. Поэтому выявление однородных производственных партий из сборных партий

изделий является одним из важнейших мероприятий при проведении испытаний с целью недопущения ошибок в оценке качества, что напрямую влияет на срок функционирования всей системы, состоящей из данных изделий. Необходимо знать, из какого количества однородных групп (кластеров) собрана производственная партия промышленной продукции.

Повысить точность методов автоматической группировки с повышенными требованиями к точности и стабильности результата позволяют предложенные в главах 2 и 3 алгоритмы, которые могут стать основой автоматизированной системы по выявлению различных по параметрам групп любых промышленных изделий.

В работе Шкабериной Г.Ш. и др. были исследованы возможности применения факторного анализа для снижения размерности данных в задаче разделения сборной партии промышленной продукции, состоящей из произвольного числа однородных партий. Не удалось выделить универсальный набор факторов для разделения сборной партии, состоящей из произвольного числа однородных партий. Таким образом, несмотря на то, что методы факторного анализа позволяют несколько сократить размерность данных, все же использование массива данных достаточно большой размерности является необходимым при применении методов кластерного анализа для разделения сборной партии (данные остаются многомерными).

В ансамблевом подходе для каждого полученного отдельными алгоритмами разбиения объектов на группы составляется предварительная бинарная матрица различий размера n x n (где n – количество объектов): $H_i=<h_i(i,j)>$, где $h_i(i,j)$ равен нулю, если элемент i и элемент j попали в один кластер, и 1, если нет.

Следующим шагом в составлении ансамбля алгоритмов кластеризации является составление согласованной матрицы бинарных разбиений.

$$H^* = \left\langle h^*(i,j) \right\rangle, \qquad h^*(i,j) = \sum w_i h_i(i,j),$$

где w_i – вес алгоритма. Мы принимаем вес, равный усредненной точности алгоритма, примененного на тестовых задачах.

Мы применили генетический алгоритм метода жадных эвристик для формирования ансамбля произвольных алгоритмов.

Точность отдельных алгоритмов кластеризации и их ансамблей можно оценить по имеющейся размеченной выборке – то есть требуется выборка, в которой принадлежность объектов к фактическим группам известна заранее.

Точность алгоритмов и их ансамблей будем оценивать следующим образом:

$$Fit^1 = A/N \to \max, \qquad (4)$$

где A – количество правильно кластеризованных объектов; N – общее количество объектов.

Алгоритм 8 Процедура составления оптимальных ансамблей алгоритмов автоматической группировки с комбинированным применением генетического алгоритма метода жадных эвристик и согласованной матрицы бинарных разбиений для практических задач

Дано: набор m тестовых задач с маркированными данными (фактическая разбивка данных на группы заранее известна), набор n алгоритмов кластеризации C_i, размер популяции q, количество алгоритмов в ансамбле p.

Решения («особи») в алгоритме – подмножества S выбранных для составления ансамбля алгоритмов кластеризации мощности p.

Шаг 1. Сгенерировать случайным образом q начальных решений – «особей» алгоритма.

Шаг 2. Для каждой особи оценить значение критерия (4), усреднённого по m задачам, применив к каждой задаче ансамбль, представленный «особью» – множеством

алгоритмов. Сохранить значение усредненного критерия в переменной Fit_j, где j – номер «особи».

Шаг 3. Проверить условия останова (превышение лимита времени), ОСТАНОВ при достижении условий.

Шаг 4. Выбрать случайным образом с равной вероятностью два номера «особей» i, j. Составить ансамбль: $S = S_i \cup S_j$.

Шаг 5. Пока $|S| > p$ выполнять:

Шаг 5.1. Для каждого i: $C_i \in S$ выполнять:

Шаг 5.1.1. Исключить i-й алгоритм из ансамбля S: $S' = S \setminus C_i$.

Шаг 5.1.2. Для S' оценить значение критерия (4), усреднённого по m задачам, применив к каждой задаче ансамбль S'. Сохранить значение усреднённого критерия в переменной Fit'_i.

Шаг 5.1.3. Перейти к следующей итерации цикла 5.1.

Шаг 5.2. Удалить из S алгоритм C_i, которому соответствует наименьшее значение Fit'_i. $S' = S \setminus C_i$.

Шаг 5.3. Следующая итерация цикла 5.

Шаг 6. Для S оценить значение критерия (4), усреднённого по m задачам, применив к каждой задаче ансамбль S. Сохранить значение усреднённого критерия в переменной Fit_{new}.

Шаг 6. Выбрать номер «особи» k с наименьшим значением Fit_k. Если $Fit_{new} > Fit_k$, то заменить k-ю особь на S. $S_k = S$; $Fit_k = Fit_{new}$.

Перейти к Шагу 2.

Данную процедуру мы применяем к описанной выше задаче составления оптимальных ансамблей алгоритмов автоматической группировки для разделения электрорадиоизделий по производственным партиям. Генетические алгоритмы метода жадных эвристик не требуют большой популяции для своей работы. Мы использовали q=10 для составления ансамблей из 3 и 5 алгоритмов (p=3, p=5).

В качестве тестовых наборов данных были проанализированы результаты неразрушающих тестовых испытаний сборных производственных партий, проведенных в специализированном тестовом центре. Сборные партии искусственно комплектовались из нескольких заведомо однородных партий. В качестве задачи ставилось разделение составленной сборной партии на однородные компоненты с последующим анализом качества этого разделения. На выходе процесса оцениваем кластеризацию по параметру точности. Под точностью мы понимаем долю объектов данных, отнесенных к «правильному» кластеру. Эту «правильность» можно оценить, имея выборку размеченных данных, для которых заранее известно их отнесение к тому или иному кластеру (Таблица 7).

Составим ансамбли из трёх и пяти соответственно лучших по точности алгоритмов кластеризации (Таблица 8) для каждого набора данных (Таблица 7).

По результатам вычислительных экспериментов видно, что любые алгоритмы автоматической группировки для задачи разделения сборной партии электрорадиоизделий или набора данных из репозитория на две однородные партии показывают довольно высокую точность. При увеличении числа однородных производственных партий в сборной партии точность падает. При этом для разных наборов данных наилучшие результаты демонстрируются разными алгоритмами.

Таблица 7 – Результаты вычислительных экспериментов над производственными партиями электрорадиоизделий отдельными алгоритмами автоматической группировки

Алгоритм	Точность / значение оптимизируемого параметра			
	140УД25АВК 2 партии	3ОТ122А 2 партии	1526LE5 6 партий	1526LE10 7 партий
k-Means-1	100,00 (Euclidean distance)	76,53 (Euclidean distance)	50,57 (Euclidean distance)	39,89 (Euclidean distance)
k-Means (fast)-1	100,00 (Euclidean distance)	67,67 (Euclidean distance)	50,57 (Euclidean distance)	39,89 (Euclidean distance)
k-Means (kernel)-1	100,00 (radial kernel)	59,19 (radial kernel)	47,14 (radial kernel)	46,83 (radial kernel)
k-Medoids-1	100,00 (Euclidean distance)	60,63 (Euclidean distance)	48,60 (Euclidean distance)	37,73 (Euclidean distance)
EM-1	96,43	90,09	нет результата	нет результата
k-Means-2	100,00 (Euclidean distance)	76,53 (Euclidean distance)	63,03 (Overlap Similarity)	52,83 (Overlap Similarity)
k-Means (fast)-2	100,00 (Euclidean distance)	76,53 (Euclidean distance)	50,99 (Kernel Euclidean distance)	46,84 (Correlation similarity)
k-Means (kernel)-2	53,57 (dot kernel)	67,67 (dot kernel)	30,22 (dot kernel)	46,83 (dot kernel)
k-Medoids-2	100,00 (Euclidean distance)	91,79 (Euclidean distance)	55,97 (Manhattan distance)	46,83 (Dice Similarity)
EM-2	96,43	95,44	нет результата	нет результата

Таблица 8 – Результаты вычислительных экспериментов с составленными ансамблями алгоритмов кластеризации

Производственная партия / ансамбль	140УД25АВК 2 партии	3ОТ122А 2 партии	1526LE5 6 партий	1526LE10 7 партий
Ансамбль из трёх	100,00	95,04	57,01	49,09
Ансамбль из пяти	100,00	95,44	52,54	47,53

Использование ансамблевого подхода более эффективно в сравнении с отдельными алгоритмами кластеризации. При этом отдельные алгоритмы способны показывать результаты, превосходящие по точности результаты ансамбля, но точность ансамбля все же выше, чем усреднённая точность отдельных алгоритмов. Так же необходимо для конкретной задачи учитывать количество алгоритмов применяемых в ансамбле, в связи с тем, что точность ансамбля алгоритмов автоматической группировки для разных наборов данных меняется при изменении числа алгоритмов в ансамбле. Поскольку на практике точность кластеризации определить невозможно вследствие отсутствия информации о фактическом составе выборки, и невозможно априорно предсказать, какой из алгоритмов в конкретном случае покажет наиболее адекватные результаты, использование ансамблевого подхода к решению подобных задач является перспективным и актуальным. В частности, применение ансамблевого подхода в сочетании с новыми алгоритмами автоматической группировки GH-VNS, обеспечивающими наилучший

результат в рамках заданной модели позволит получать результаты не только более адекватные, но и воспроизводимые при многократных запусках алгоритма.

ЗАКЛЮЧЕНИЕ

В диссертации предложены новые алгоритмы метода жадных эвристик (в том числе параллельные) для решения задач автоматической группировки (кластеризации) объектов, сочетающие применение жадных агломеративных эвристических процедур и расширенный локальный поиск с чередующимися рандомизированными окрестностями, позволяющие решать круг практических задач с повышенной точностью результата (по достигаемому значению целевой функции), а также процедура составления ансамблей алгоритмов автоматической группировки.

Цель диссертации достигается путем решения поставленных задач, а именно:

1. Анализ существующих проблем при применении методов автоматической группировки объектов, к которым предъявляются высокие требования по точности и стабильности результата, выявил дефицит алгоритмов, способных выдавать за фиксированное время результаты, которые было бы трудно улучшить известными методами, и которые бы обеспечивали стабильность получаемых результатов при многократных запусках алгоритма. При этом известные алгоритмы метода жадных эвристик требуют значительных вычислительных затрат.

2. Разработаны новые алгоритмы автоматической группировки объектов в соответствии с оптимизационной моделью k-средних, основанные на совместном применении алгоритма k-средних, жадных агломеративных эвристических процедур и расширенного локального поиска с чередующимися рандомизированными окрестностями. При этом вид окрестности поиска определяется видом применяемой жадной агломеративной эвристической процедуры, а случайным образом генерируемое известное решение является параметром данной окрестности. Показано, что новые алгоритмы позволяют получать более точный и стабильный результат (по достигаемому значению целевой функции) в сравнении с известными алгоритмами, являясь конкурентоспособными в сравнении с известными алгоритмами метода жадных эвристик при фиксированном лимите времени работы алгоритма, позволяющем использовать алгоритмы в интерактивном режиме принятия решений.

3. Разработаны новые алгоритмы автоматической группировки объектов, основанной на модели k-медоид, также основанные на совместном применении жадных агломеративных эвристических процедур, расширенного локального поиска с чередующимися рандомизированными окрестностями и алгоритма Partition around Medoids. Показано, что новые алгоритмы также позволяют получать более точный и стабильный результат (по достигаемому значению целевой функции) в сравнении с известными алгоритмами.

4. Разработаны новые алгоритмы четкой кластеризации объектов, основанной на модели разделения смеси вероятностных распределений с применением жадных агломеративных эвристических процедур, расширенного локального поиска с чередующимися рандомизированными окрестностями и известного классификационного EM-алгоритма, также обладающие преимуществами по получаемому значению целевой функции за фиксированное время. Это позволяет говорить о новом подходе к разработке эффективных алгоритмов автоматической группировки, основанном на комбинированном применении известных для соответствующих задач алгоритмов локального поиска, жадных агломеративных эвристических процедур и алгоритмов поиска с чередующимися рандомизированными окрестностями, образуемыми применением одной из жадных агломеративных эвристических процедур к лучшему известному решению и второму решению, генерируемому случайным образом и являющемуся параметром окрестности.

5. Впервые предложены параллельные модификации алгоритмов метода жадных эвристик для архитектуры CUDA, позволяющие существенно расширить рамки применения метода жадных эвристик и охватить достаточно большие задачи – до сотен тысяч векторов многомерных данных.

6. Разработана процедура составления оптимальных ансамблей алгоритмов автоматической группировки с комбинированным применением генетического алгоритма метода жадных эвристик и согласованной матрицы бинарных разбиений для практических задач, позволяющая уменьшить число ошибок при разделении сборной партии промышленной продукции на однородные партии с использованием данных неразрушающих тестовых испытаний.

ПУБЛИКАЦИИ ПО ТЕМЕ ДИССЕРТАЦИИ

Публикации в изданиях, рекомендованных ВАК России:

1. Рожнов, И.П. Дальнейшее развитие метода жадных эвристик для задач автоматической группировки объектов / Л.А.Казаковцев, Д.В.Сташков, И.П.Рожнов, О.Б.Казаковцева // Системы управления и информационные технологии.– 2017.– № 4(70). С. 34-40.

2. Рожнов, И.П. Анализ алгоритмов кластеризации и их ансамблей для задачи выделения производственных партий электрорадиоизделий / В.И.Орлов, И.П.Рожнов, О.Б.Казаковцева, Л.А.Казаковцев // Экономика и менеджмент систем управления.– 2017.– № 4.4 (26). С. 486-492.

3. Рожнов, И.П. Усовершенствованная методика формирования партий электронной компонентной базы с особыми требованиями качества / В.И.Орлов, Д.В.Сташков, Л.А.Казаковцев, И.П.Рожнов, О.Б.Казаковцева, И.Р.Насыров // Современные наукоемкие технологии.– 2018.– № 1. С. 37-42.

4. Рожнов, И.П. 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.

5. Рожнов, И.П. Алгоритм для задачи k-средних с рандомизированными чередующимися окрестностями / И.П.Рожнов, Л.А.Казаковцев, М.Н.Гудыма, В.Л.Казаковцев // Системы управления и информационные технологии.– 2018.– № 3 (73). С. 46-51.

6. 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.

7. Рожнов, И.П. Реализация жадных эвристических алгоритмов кластеризации для массивно-параллельных систем / И.П.Рожнов, В.Л.Казаковцев // Системы управления и информационные технологии.– 2019.– № 2 (76). С. 36-40.

Публикации в изданиях, входящих в системы цитирования WoS и Scopus:

8. Rozhnov I. Ensembles of clustering algorithms for problem of detection of homogeneous production batches of semiconductor devices / Rozhnov I. Orlov V., Kazakovtsev L. // В сборнике: CEUR Workshop Proceedings Сер. "OPTA-SCL 2018 – Proceedings of the School-Seminar on Optimization Problems and their Applications". CEUR-WS. 2018. Vol.2098. P.338-348.

9. Rozhnov I.P. Increase in Accuracy of the Solution of the Problem of Identification of Production Batches of Semiconductor Devices / Rozhnov I.P., Orlov V.I., Kazakovtsev L.A. // 14th International Scientific-Technical Conference on Actual Problems of Electronic Instrument Engineering, APEIE 2018. P.363-367. DOI:10.1109/APEIE.2018.8546294.

10. Rozhnov I.P. Variable neighbourhood search algorithm for k-means clustering / Orlov V I, Kazakovtsev L A, Rozhnov I P, Popov N A and Fedosov V V // IOP Conf. Series: Materials

Science and Engineering 2018. Vol. 450, Article ID 022035, DOI:10.1088/1757-899X/450/2/022035.

11. Rozhnov I.P. Parallel implementation of the greedy heuristic clustering algorithms / Kazakovtsev L A, Rozhnov I P, Popov E A, Karaseva M V and Stupina A A // IOP Conf. Series: MIP: Engineering-2019. Vol. 537. DOI:10.1088/1757-899X/537/2/022052.

12. Rozhnov I. Improved Classification EM algorithm for the Problem of Separating Semiconductor Device Production Batches / Rozhnov I, Kazakovtsev L, Bezhitskaya E and Bezhitskiy S // IOP Conf. Series: MIP: Engineering-2019. Vol. 537. DOI:10.1088/1757-899X/537/5/052032.

В других изданиях:

13. Рожнов, И.П. Алгоритм для серии задач разделения смеси распределений / Д.В.Сташков, М.Н.Гудыма, Л.А.Казаковцев, И.П.Рожнов, В.И.Орлов // Решетневские чтения: Материалы XXI междунар. науч.-практ. конф. в 2-х частях. Красноярск: СибГУ, 2017. – Ч. 1 – С. 327-328.

14. Рожнов, И.П. Усовершенствованный алгоритм разделения смеси распределений для данных большой размерности / Л.А.Казаковцев, Д.В.Сташков, О.Б.Казаковцева, И.П.Рожнов, А.В.Медведев // Лесной и химический комплексы – проблемы и решения: Материалы всероссийской науч.-практ. конф. Красноярск: СибГУ, 2017. – С. 502-505.

15. Рожнов, И.П. Выделение партий электрорадиоизделий ансамблями алгоритмов кластеризации / И.П.Рожнов, Л.А.Казаковцев, В.И.Орлов // В книге: Проблемы оптимизации и их приложения. Тезисы докладов VII Международной конференции: памяти профессора А.А. Колоколова.– 2018.– С. 90.

16. Рожнов, И.П. Алгоритм поиска в чередующихся окрестностях для задачи выделения однородных производственных партий электрорадиоизделий / В.И.Орлов, И.П.Рожнов, В.Л.Казаковцев, М.Н.Гудыма // Решетневские чтения: Материалы XXII междунар. науч.-практ. конф. в 2-х частях. Красноярск: СибГУ, 2018. – Ч. 1 – С. 315-316.

17. Рожнов, И.П. Формирование электронной компонентной базы с особыми требованиями качества с применением ансамблей алгоритмов кластеризации / И.П.Рожнов, В.И.Орлов, Л.А.Казаковцев // Решетневские чтения: Материалы XXII междунар. науч.-практ. конф. в 2-х частях. Красноярск: СибГУ, 2018. – Ч. 1 – С. 322-324.

18. Рожнов, И.П. Усовершенствованный СЕМ-алгоритм для данных большой размерности / Л.А.Казаковцев, И.П.Рожнов, П.Ф.Шестаков // Наука и образование: опыт, проблемы, перспективы развития: Материалы междунар. науч.-практ. конф. Красноярск: КрасГАУ, 2019. – С. 244-247.

Свидетельство о государственной регистрации программы для ЭВМ:

19. Рожнов, И.П. Система составления оптимальных ансамблей алгоритмов кластеризации для задачи выделения производственных партий электрорадиоизделий / В.И.Орлов, И.П.Рожнов, Л.А.Казаковцев, С.М.Голованов – М.: РОСПАТЕНТ. 2019. Свидетельство № 2019610095 от 09.01.2019.