На правах рукописи

Сташков Дмитрий Викторович

СИСТЕМЫ АВТОМАТИЧЕСКОЙ ГРУППИРОВКИ ОБЪЕКТОВ НА ОСНОВЕ РАЗДЕЛЕНИЯ СМЕСИ РАСПРЕДЕЛЕНИЙ

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 Сибирском государственном университете науки и технологий им. академика М.Ф. Решетнева.

Научный руководитель: доктор технических наук, доцент **Казаковцев Лев Александрович**

Официальные оппоненты:

Кельманов Александр Васильевич
доктор физико-математических наук, старший научный сотрудник, главный научный сотрудник лаборатории анализа данных Федерального государственного бюджетного учреждения науки «Институт математики им. С. Л. Соболева» СО РАН (г. Новосибирск)

Крутиков Владимир Николаевич
доктор технических наук, доцент, профессор кафедры прикладной математики федерального государственного бюджетного образовательного учреждения высшего образования «Кемеровский государственный университет»

Ведущая организация: Федеральное государственное бюджетное образовательное учреждение высшего образования "Воронежский государственный технический университет"

Защита состоится 8 декабря 2017 г. в 13 часов на заседании диссертационного совета Д 212.249.05 Сибирского государственного университета науки и технологии 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 технологии им. академика М.Ф. Решетнева.

Автореферат разослан “____” _____ 2017 г.

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

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность. Применение систем автоматической группировки (кластеризации) находит все большее распространение в связи с расширением областей применения задач анализа данных, таких как распознавание изображений, решение задач диагностирования в медицине, маркетинговые исследования, исследование трафика Интернет и пр.

В данном исследовании предполагается, что совокупность измерений (параметров) объектов одной группы (кластера) является многомерной случайной величиной, распределенной по некоторому известному закону распределения, при этом параметры этого распределения неизвестны. Суть рассматриваемой задачи группировки объектов состоит в том, чтобы отделить объекты, предположительно порожденные одним из распределений в этой смеси, от других. Иными словами, задача состоит в поиске смеси распределений из некоторого допустимого класса смесей. При этом класс допустимых смесей определяется классом допустимых смешиваемых распределений. Традиционным инструментом при решении задач разделения смесей является метод максимального правдоподобия и соответствующий критерий – функция правдоподобия. Задача разделения смесей сводится к задаче максимизации этой функции. При этом искомыми параметрами функции правдоподобия являются параметры распределений.

Модели, основанные на разделении смесей вероятностных распределений, используемые, кроме кластерного анализа, также при непараметрическом оценивании плотности, демонстрируют высокую адекватность при описании неоднородных данных.

Наиболее популярным методом численного решения оптимизационной задачи разделения смеси распределений является EM-алгоритм (Expectation Maximization – максимизация математического ожидания) и его модификации, такие как SEM (Stochastic EM – стохастический EM). Алгоритм имеет ряд существенных недостатков. В частности, он неустойчив по отношению к исходным данным. Кроме того, алгоритм требует задания предполагаемого числа распределений в смеси. Алгоритм и его модификации требуют задания начального решения, являясь (не в строгом смысле) алгоритмами локального поиска. Алгоритм неустойчив к выбору начального решения.

Как показывают вычислительные эксперименты, стабильность результатов при многократных запусках SEM-алгоритма выше, чем у EM-алгоритма. Хотя истинный оптимум для больших задач определить в общем случае практически невозможно, об имеющемся резерве в плане улучшения результатов говорит то, что при длительных вычислительных экспериментах результаты лучших попыток выполнения EM-алгоритма и его модификаций могут различаться на десятки процентов по значению целевой функции правдоподобия от усредненных значений всего множества попыток.

В некоторых случаях решение задач автоматической группировки требует получения результата в рамках предложенной постановки задачи, который было бы крайне трудно существенно улучшить известными методами. Под улучшением результата понимается увеличение достигнутого значения целевой функции правдоподобия. Такие задачи возникают, если цена ошибки очень высока. Кроме того, при решении таких задач могут затрагиваться интересы различных сторон, и решение задачи одной стороной должно гарантировать, что другая сторона не поставит оптимальность результата под сомнение, предъявив иной результат в рамках

предложенной модели группировки. Данная работа посвящена разработке подхода к построению систем автоматической группировки, обеспечивающих данное свойство результата,на примере системы группировки электрорадиоизделий (ЭРИ) по однородным производственным партиям.

Степень разработанности темы исследования. Исследование свойств смесей вероятностных распределений в целях моделирования новых распределений было начато в 1880-е годы С.Ньюкомбом и К.Пирсоном, в дальнейшем было продолжено в 1980-е годы Б.Эвериттом, Д.М.Титтерингтоном, Дж.Маклаланом и др.

Ю.И. Журавлёвым был введен и изучен новый класс алгоритмов распознавания (классификации) - алгоритмы вычисления оценок, которые служат универсальным языком описания процедур распознавания, являющихся составной частью алгоритмов автоматической группировки (кластеризации) на основе разделения смеси распределений, таких как ЕМ-алгоритм. Систематизация постановок задач кластеризации и алгоритмов их решения, включая ЕМ-алгоритм, а также формулировка общей экстремальной задачи кластеризации выполнена С.А. Айвазяном и др. Отдельным направлением можно считать развитие методов бикластеризации (разбиения на две группы), которым посвящены работы А.В. Кельманова. Н.Г. Загоруйко и др. разработаны методы автоматической классификации (таксономии), выбора информативных признаков, построения решающих функций, обнаружения ошибок, которые нашли применение в геологии, медицине, генетике, экономике и во многих других прикладных областях.

Процедура, схожая с ЕМ-алгоритмом, предложена в 1926 г. (A.G. McKendrick) и получила развитие в работах 1950-70 годов (M.J.R. Healey, M.H. Westmacott, М.И. Шлезингером и др.). Название «ЕМ-алгоритм» предложено в 1977 (A. Dempster и др.). В работах В.Ю. Королева и его последователей (напр., А.Л. Назаров, А.К. Горшенин) проведена систематизация подходов к численному решению задач, предложены модификации наиболее популярного ЕМ-алгоритма с повышенной стабильностью результата с использованием медианных модификаций ЕМ-алгоритма, исследованы свойства устойчивости этих модификаций к возмущениям в данных, доказана сходимость SEM-алгоритма к стационарному распределению, построены критерии определения числа компонент смеси, а также предложены так называемые сеточные алгоритмы разделения смеси, вычислительная сложность которых довольно высока. В 2006 г. предложен подход, основанный на комбинации принципа максимизации функции правдоподобия и алгоритма *k*-средних, эффективный на малых объемах данных (S. Nasser, R. Alkhaldi, G. Vert). В работах И. Мелийсона показано, что метод Ньютона или другие градиентные методы являются более быстрыми и обеспечивают нужную точность оценивания, но при этом проявляют нестабильность и требуют аналитической обработки функции правдоподобия. Также унифицирован ЕМ-подход и метод Ньютона.Ч.Лиу, Д.Б.Рубин и др. предложен метод, позволяющий улучшить скорость работы ЕМ-алгоритма через “настройку ковариации” за счет получения дополнительной информации, полученной в оценке полных данных.

Тема локального поиска с чередующимися окрестностями, весьма эффективного для многих многоэкстремальных NP-трудных задач, раскрыта в работах Ю. А. Кочетова, Н. Младеновича (N.Mladenovic), П. Хансена (P.Hansen). Авторами исследованы границы возможностей локального поиска. Гибкость и легкость адаптации этой идеи к различным моделям, объясняют ее

конкурентоспособность при решении NP-трудных задач,в частности задач о p-медиане (T.G.Crainic, M.Gendreau, N.Hoebetal., 2001), задачи Вебера (J.Brimberg, P.Hansen, N.Mladenovic, E.Taillard, 2000), задачкластеризации и размещения (J.Brimberg, N.Mladenovic et al., 1996) и других. Согласно «гипотезе о большой долине», локальные минимумы распределены неравномерно и расположены достаточно близко друг к другу, занимая малую часть допустимой области (K. D.Boese, A. B.Kahng, S.Muddu, 1994). Она позволяет сузить область поиска глобального оптимума, продолжая поиск близко к обнаруженным локальным оптимумам. Именно эта гипотеза лежит в основеразличных операторов скрещивания (crossover)для генетических алгоритмов (D. E.Goldberg, 1989) и многих других подходов.

Метод жадных эвристик предложен в 2014 году Л.А.Казаковцевым и А.Н.Антамошкиным[1,2] для задач автоматической группировки на основе моделей теории размещения. Метод широко использует эволюционные подходы, большой вклад в развитие которых вносит красноярская школа эволюционных методов оптимизации (Е.С.Семенкин и др.). Метод представляет собой расширяемый подход для построения систем размещения, кластеризации, псевдобулевой оптимизации. Построенные системы дают хорошие результаты (по значению целевой функции и стабильности этих значений) для задач автоматической группировки с большим количеством объектов – до сотен тысяч, представленных векторами данных большой размерности – до сотен измерений.

Идея настоящего исследования состоит в применении к задачам автоматической группировки на основе разделения смесей распределений подходов, на которых основан метод жадных эвристик. Данный метод обеспечивает наилучшие и наиболее стабильные значения целевых функций, используемых в задачах автоматической группировки, основанных на моделях теории размещения. В работе выдвигается и эмпирически доказывается гипотеза о том, что подходы, на которых основан метод жадных эвристик, а также подходы с использованием поиска в чередующихся окрестностях, позволяют для задач разделения смесей получить аналогичный результат: улучшить получаемые значения целевой функции, а также стабильность этих значений при многократных запусках алгоритма за ограниченное время.

Объектом диссертационного исследования являются задачи автоматической группировки на основе разделения смеси вероятностных распределений, **предметом исследования** – алгоритмы их решения.

Целью исследования является повышение точности результата решения задачи разделения смеси распределений, выраженного значением целевой функции правдоподобия, за ограниченное время.

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

1. Провести анализ проблем, возникающих при применении методов кластеризации, основанных на модели разделения смеси распределений, а также анализ методов, позволяющих повысить точность и стабильность результата схожих по своей постановке задач;

---

[1] Казаковцев Л.А. Метод жадных эвристик для задач размещения / Л.А.Казаковцев, А.Н.Антамошкин // Вестник СибГАУ.–2015.–№2.–С.317-325.

[2] Kazakovtsev L.A. Genetic Algorithm with Fast Greedy Heuristic for Clustering and Location Problems / L. A. Kazakovtsev, A.N. Antamoshkin. // Informatica.- 2014.- Vol. 38, No. 3.- P.229-240.

2. Построить более эффективные (по сравнению с известными) алгоритмы решения задач автоматической группировки объектов на основе разделения смеси распределений в пространствах большой размерности (до сотен измерений) с применением жадных агломеративных эвристических процедур и идей метода поиска с чередующимися окрестностями. Под эффективностью здесь понимается способность алгоритма получать результат с наилучшим значением целевой функции правдоподобия за ограниченное время, приемлемое для работы алгоритма в интерактивном режиме;

3. Построить эффективные генетические алгоритмы метода жадных эвристик, обеспечивающие наилучшиезначения целевой функции правдоподобия для больших задач автоматической группировки на основе разделения смеси распределений за ограниченное время;

4. Разработать подход к составлению эффективных комбинаций алгоритмов для систем автоматической группировки объектов на основе разделения смеси распределений.

5. Построить модель разделения сборных партий промышленной продукции на однородные партии на основе модели разделения смеси распределений, исследовать ее адекватность на практических примерах.

Методы исследования. Методологической базой явились работы по методам кластеризации, в частности – по методам решения задач разделения смеси распределений с применением EM-алгоритма и его модификаций, а также по методу жадных эвристик для задач автоматической группировки объектов. Использован математический аппарат теории вероятностей, теории размещения, методы теории оптимизации, в частности – эволюционные методы оптимизации, а также методы системного анализа, исследования операций.

Новые научные результаты и положения,выносимые на защиту:

1. Разработаны новые генетические алгоритмы, основанные на идеях метода жадных эвристик для задач разделения смеси распределений. Алгоритмы позволяют получать более точный по значению целевой функции результат по сравнению с известными методами для задач с наборами данных большой размерности (сотни измерений).

2. Разработаны новые алгоритмы поиска с чередующимися окрестностями для задач разделения смеси распределений. Алгоритмы позволяют получать более точный по значению целевой функции результат по сравнению с известными методами за ограниченное время для больших задач (до сотен тысяч векторов данных) с наборами данных большой размерности (сотни измерений). Новые алгоритмы расширяют круг возможных стратегий поиска, применяемых в составе метода жадных эвристик.

3. Представлена новая модель разделения сборных партий промышленных изделий по результатам множественных тестовых испытаний на основе разделения смеси сферических или некоррелированных гауссовых распределений. Модель позволяет снизить процент ошибочно сгруппированных элементов сборной партии по сравнению с известными моделями.

Значение для теории. Результаты исследования дополняют арсенал эффективных эвристических методов решения задач автоматической группировки объектов на основе разделения смеси вероятностных распределений. Кроме того, расширено множество возможных стратегий поиска, применяемых в составе метода

жадных эвристик, что создает фундамент для разработки новых алгоритмов для других оптимизационных задач.

Практическая ценность Разработанные алгоритмы позволяют повысить точность решений систем автоматической группировки объектов на основе модели разделения смеси вероятностных распределений за ограниченное время. Новая модель разделения сборных партий промышленной продукции на однородные партии позволяет снизить процент ошибок при выявлении однородных партий.

Практическая реализация результатов: Программная реализация алгоритма решения задачи автоматической классификации ЭРИ по однородным производственным партиям с использованием данных неразрушающих тестовых испытаний в составе СППР автоматической классификации ЭРИ по производственным партиям ОАО ИТЦ – НПО ПМ (г.Железногорск) обогатило данную СППР возможностью применения дополнительной модели группировки изделий, служащей в настоящий момент в качестве средства верификации основной модели группировки, основанной на модели k-средних с особым способом нормирования данных.

Апробация работы. Основные положения и результаты диссертационной работы докладывались и обсуждались на международных конференциях. В их числе: XVI Международная научно-практическая конференция «Приоритетные научные направления: от теории к практике» (2015 г., г. Новосибирск),XX Международная научная конференция «Решетневские чтения» (2016 г., г. Красноярск), 55-я международная студенческая конференция «МНСК-2017» (2017 г., г. Новосибирск),7-я международная конференция «Системный анализ и информационные технологии» (2017 г., г.Светлогорск), XVII Байкальская международная школа-семинар «Методы оптимизации и их приложения» (2017 г., с. Максимиха, Бурятия). Диссертация в целом обсуждалась на семинаре «Алгоритмы автоматической группировки объектов» в ИМ СО РАН (г.Новосибирск, 24.04.2017 г.).

Публикации. Основные теоретические и практические результаты диссертации опубликованы в 14 изданиях (среди них 1 рецензируемая монография, также имеется 1 свидетельство о государственной регистрации программы для ЭВМ), среди которых 5 работ в ведущих российских рецензируемых изданиях, рекомендуемых в действующем перечне ВАК, 1 - в международном издании, проиндексированном в системе цитирования Scopus.

Структура и объем диссертации. Диссертация состоит из введения, 3 глав и заключения. Она изложена на 172 листах машинописного текста, содержит список литературы из 208 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.

В процессе группировки выполняется выделение подмножеств (кластеров) данных на основании установленного отношения схожести элементов. Наиболее простые и в тоже время эффективные методы и модели основываются на минимизации суммарных расстояний между объектами одной группы (кластера) или между объектами кластера и его центром или, иначе - центроидом, медоидом,

медианой в зависимости от постановки конкретной задачи (задачи *k*-средних, *k*-медоид или дискретная *p*-медианная, непрерывная *p*-медианная).

Более сложные, но в тоже время более гибкие и информативные модели основаны на разделении смесей вероятностных распределений. Предполагается, что выборка содержит объекты разных групп, то есть параметры объектов порождены различными распределениями.

EM-алгоритм для разделения смеси *k* нормальных распределений может быть описан следующим образом (на примере сферических гауссовых распределений). Алгоритм стартует с начальными стартовыми значениями параметров $\mu_i\langle 0\rangle$, $\alpha_i\langle 0\rangle$, $\sigma_i\langle 0\rangle$ (математическое ожидание, априорная вероятность и среднеквадратичное отклонение *i*-го распределения соответственно), которые в дальнейшем обновляются в соответствии со следующей двухшаговой процедурой (здесь *t* – номер итерации).

Алгоритм 1. EM-алгоритм (на примере разделения сферических гауссовых распределений).

Шаг 1 (E-шаг). Пусть $\tau_i \sim N(\mu_i^{\langle t\rangle}, \sigma_i^{\langle t\rangle 2} I_n)$ является плотностью *i*-го гауссова распределения: $\tau_i(x) = (2\pi)^{-n/2}\sigma_i^{-n}\exp(\|x-\mu_i\|^2 / 2\sigma_i^{\,2})$. Для каждого вектора исходных данных $x \in R^n$ и каждого $1 \le i \le k$ по формуле Байеса вычислим условную вероятность того, что *x* относится к *i*-му распределению с учетом текущих параметров: $p_i^{\langle t+1\rangle}(x) = \alpha_i^{\langle t\rangle}\tau_i(x) / (\sum_j \alpha_j^{\langle t\rangle}\tau_j(x))$.

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

$$\sigma_i^{\langle t+1\rangle 2} = \frac{1}{d}\sum_{x\in S}\left\|x-\mu_i^{\langle t+1\rangle}\right\|^2 p_i^{\langle t+1\rangle}(x) \qquad (1)$$

Повторять с шага 1.

Останов алгоритма, если нет приращения целевой функции правдоподобия

$$L^{<t+1>} = \sum_{x\in S}\sum_{i=1}^{k}\ln(\tau_i p_i^{\langle t+1\rangle}(x)). \qquad (2)$$

В случае некоррелированных гауссовых распределений алгоритм оперирует вектором среднеквадратичных отклонений для каждого кластера. Использовался подход с разделением смесей некоррелированных гауссовых распределений с одинаковыми для всех кластеров векторами среднеквадратичных отклонений. В этом случае σ_j – среднеквадратичное отклонение (одинаковое для всех кластеров) по *j*-му измерению, и его перерасчет вместо (1) осуществляется следующим образом: $\sigma_j^{\langle t+1\rangle 2} = \sum_{i=1}^{k}\sum_{x\in S}\left\|x-\mu_i^{\langle t+1\rangle}\right\|^2 p_i^{\langle t+1\rangle}(x) / (N\alpha_i)$.

Соответственно, в случае многомерного гауссова распределения алгоритм оперирует полными ковариационными матрицами и обратными матрицами.

Более простые модели автоматической группировки, основанные на минимизации расстояний между объектами или суммарных расстояний от объектов до специальных новых объектов – центров кластеров, достаточно просты, и для них разработаны быстрые эффективные алгоритмы.

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: $\arg\min F(X_1,...,X_k) = \sum_{i=1}^{N} \min_{j\in\{1,k\}} L(X_j, A_i)$. Здесь $L()$ – некая функция расстояния между двумя точками в *d*-мерном пространстве. При $L(X, A_i) = \|X - A_i\|^2$ имеем задачу *k*-средних.

Для решения задачи *k*-средних может быть использован одноименный алгоритм, называемый также ALA-алгоритмом (Alternating Location-Allocation – чередующееся размещение-распределение), включающий два чередующихся шага:

Алгоритм 2. 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.

Этот же алгоритм может эффективно использоваться и для непрерывных *p*-медианных задач.

За исключением особых случаев, задачи *k*-средних и *p*-медианная, как и задача разделения смеси распределений, являются NP-трудными, требующими глобального поиска. Результат ALA-процедуры зависит от выбора начальных центров кластеров, это является проблемой с точки зрения обеспечения воспроизводимости результатов вычислений.

Аналогично, выбор начальных $\mu_i\langle 0\rangle$, $\alpha_i\langle 0\rangle$, $\sigma_i\langle 0\rangle$ определяет результат EM-алгоритма. В несколько меньшей степени зависимость от начального решения прослеживается и для других модификаций EM-алгоритма (например, SEM).

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.

Метод жадных эвристик для задач автоматической группировки предусматривает применение широкого круга стратегий глобального поиска. Простейшей стратегией является мультистарт жадных агломеративных эвристических процедур. При этом предусматривается применение различных алгоритмов локального поиска, характерных и наиболее эффективных для той или иной практической задачи. Метод жадных эвристик для задач автоматической группировки на *k* групп (кластеров) можно представить, как три вложенных цикла:

А) Цикл, осуществляющий стратегию глобального поиска, формирующий промежуточные решения, представленные множествами, мощность которых выше *k*, для которых запускаются жадные агломеративные процедуры.

Б) Цикл жадной агломеративной эвристической процедуры, приводящий мощность промежуточного решения к значению *k*, в ходе которого могут запускаться алгоритмы локального поиска, улучшающие промежуточные решения.

В) Цикл локального поиска. В алгоритмах для задач *k*-средних и аналогичных авторы предлагают использовать наиболее эффективные для таких задач двухшаговые алгоритмы: процедуру Ллойда, PAM-процедуру.

Можно отметить общность состава оптимизируемых параметров задач кластеризации на основе модели *k*-средних и аналогичных (параметрами являются координаты центров кластеров) с задачами на основе модели разделения смеси распределений, в которых параметрами являются параметры распределений (в том числе математические ожидания – фактически центры кластеров), дополненные априорными вероятностями распределений α_i. Кроме того, наиболее популярные алгоритмы – ALA и EM схожи по структуре: оба являются процедурами с двумя чередующимися шагами распределения объектов по кластерам и размещения. В случае EM-алгоритма E-шаг фактически распределяет объекты между предполагаемыми распределениями, приписывая каждому объекту значение вероятности $p_i^{\langle t+1\rangle}(x)$ того, что объект порожден *i*-м распределением, а M-шаг модифицирует параметры распределений аналогично тому, как ALA-процедура модифицирует координаты центров. Наконец, целевые функции обладают свойствами многоэкстремальности. Общность свойств моделей и алгоритмов дает обоснованную надежду на применение схожих методов повышения точности результатов решения.

В главе 2 предлагаются новые подходы к построению алгоритмов для решения задач автоматической группировки на основе разделения смеси распределений и проводится сравнительный анализ их эффективности.

Идея жадных агломеративных алгоритмов для задач автоматической группировки заключается в последовательном уменьшении числа групп (кластеров), на которые разбивается выборка. В нашем случае уменьшается число распределений (см. шаг 5 следующего алгоритма).

Алгоритм 3.Базовая жадная агломеративная эвристика для разделения распределений (на примере некоррелированных гауссовых распределений).Дано: начальное число гауссовых распределений (кластеров) *K*, требуемое число кластеров *k, K>k.*

1.Выбрать начальное решение с *K* кластерами, т.е. выбрать случайным образом начальные параметры пары множеств параметров распределений и их весовых коэффициентов $< D,W >=< \{N(\mu_i, \sigma_i^2 I_n)\}, \{\alpha_i\}, i = \overline{1,K} >$.

2.Выполнить Алгоритм 1, получить новое (улучшенное) решение задачи, представленное <*D,W*>.

3.Если *K*=*k*, то останов.

4.Для каждого $i' \in \{\overline{1,K}\}$ выполнять:

4.1. Получить пару усеченных множеств $< D', W' >=< D \setminus \{N(\mu_{i'}, \sigma_{i'}^2 I_n)\}, \quad W \setminus \{\alpha_{i'}\} >$.

4.2. Запустить Алгоритм 1 с начальными значениями параметров распределений, представленных усеченным <*D',W'*>. При этом Алгоритм 1 ограничивается одной итерацией. Для полученного EM-алгоритмом решения рассчитать целевую функцию *L* согласно (2), сохранить ее значение в переменной $L'_{i'}$. Следующая итерация цикла 4.

5. Найти индекс $i'' = \arg\max_{i'=\overline{1,k}} L_{i'}$. Получить пару усеченных $< D'', W'' >=< D \setminus \{N(\mu_{i''}, \sigma_{i''}^2 I_n)\}, \quad W \setminus \{\alpha_{i''}\} >$. Запустить для этой пары усеченных множеств Алгоритм 1, затем перейти к шагу 3.

Такой алгоритм может стартовать из случайных начальных решений в режиме мультистарта, или из начального решения, полученного объединением двух известных решений задачи:

Алгоритм 4. Жадная процедура с частичным объединением №1. Дано: пары множеств распределений и весовых коэффициентов $< D',W' >=< \{N(\mu'_i, \sigma'^2_i I_n)\}, \{\alpha'_i\}, i = \overline{1,k} >$ и $< D'',W'' >=< \{N(\mu''_i, \sigma''^2_i I_n)\}, \{\alpha''_i\}, i = \overline{1,k} >$.

1. Для каждого $i' \in \overline{1,k}$ выполнять:

1.1. Объединить множества пары<*D',W'*>с очередными элементами множеств пары<*D'',W''*>: $< D,W >=< D' \cup \{N(\mu''_{i'}, \sigma''^2_{i'} I_n)\}, W' \cup \{\alpha''_{i'}\} >$.

1.2. Запустить Алгоритм 3 (с Шага 2) с этими парами объединенных множеств <*D,W*> в качестве начального решения. Полученный результат (пару полученных множеств, а также значение целевой функции) сохранить.

2. Возвратить в качестве результата лучшее (по значению целевой функции) из решений, полученных на шаге 1.2.

Данная процедура фактически дополняет один из «родительских» вариантов решения задачи разделения смесей, представленных парой <*D',W'*> поочередно с каждым из элементов второго «родительского» решения, представленного соответствующей парой множеств <*D'',W''*>, затем к этой паре объединенных множеств применяет жадную эвристику (Алгоритм 3) и фиксирует лучший из полученных результатов.

Более простой, но более вычислительно затратный вариант:

Алгоритм 5. Жадная процедура с полным объединением.

1.Объединить множества <*D',W'*> и <*D'',W''*>:

$< D,W >=< D' \cup D'', \ W' \cup W'' >$. Запустить Алгоритм 3(с Шага 2) с этими объединенными множествами в качестве начального решения.

Как видно, здесь выполняется полное объединение двух «родительских» вариантов решения задачи. Наконец, возможен промежуточный вариант:

Алгоритм 6. Жадная процедура с частичным объединением №2.

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

2. Повторять *k-r* раз:

2.1. Сформировать случайно выбранное подмножество *D'''* элементов множества *D''* мощности *r* и подмножество *W'''* соответствующих элементов множества *W''* (также мощности *r*).Объединить множества $< D,W >=< D' \cup D''', \ W' \cup W''' >$.Запустить Алгоритм 3 с<*D,W*> (с Шага 2).

3. Возвратить в качестве результата лучшее (по значению целевой функции) из решений, полученных на шаге 2.1.

В диссертации представлены результаты работы перечисленных жадных процедур в режиме мультистарта из случайных начальных решений. Показано, что для некоторых задач они показывают лучшие результаты, чем известные методы.

Данные эвристические процедуры, являющиеся (не в строгом смысле) алгоритмами локального поиска в окрестности известного («родительского») решения, представленного множествами <*D',W'*>, могут использоваться в составе различных стратегий глобального поиска. При этом в качестве окрестностей, в которых производится поиск решения, используются решения, производные («дочерние») по отношению к решению <*D',W'*>, образованные комбинированием его

элементов с элементами решения <*D’’,W’’*> и применением жадной агломеративной эвристики (Алгоритма 3).

Метод жадных эвристик предусматривает в качестве одного из вариантов организации локального поиска применение алгоритмов поиска в чередующихся окрестностях (VNS-алгоритмы – Variable Neighborhood Search). В работах Н.Младеновича и Ю.А.Кочетова приводится обзор современных методов локального поиска, основанных на идее чередующихся окрестностей, показаны пути гибридизации этих методов с другими метаэвристиками, демонстрируются разные способы построения окрестностей и их свойства.

Суть алгоритма с чередующимися окрестностями заключается в том, что для некоторого промежуточного решения *X* есть конечное множество окрестностейS_N этого решения. Выбирается очередная окрестность *S*, в которой с применением соответствующего алгоритма локального поиска производится поиск решения с лучшим значением целевой функции. Если такое решение найдено, *X* заменяется этим новым решением, и поиск в той же окрестности *S* продолжается. Если такое решение найти не удается, из множества S_N выбирается новая окрестность, и поиск продолжается в ней.

Задача разделения смесей распределений является задачей глобальной непрерывной оптимизации. С одной стороны, алгоритм глобального поиска должен периодически «выбивать» промежуточное решение задачи из «области притяжения» локального оптимума, а сдругой стороны, решения, образованные из элементов различных локально-оптимальных решений, с большей вероятностью оказываются ближе к глобальному оптимуму, чем случайно выбранные решения, что показано, например, в работах А.В.Еремеева. Таким образом, перспективным представляется поиск в окрестности локального оптимума, образованной заменой отдельных элементов локально-оптимального решения элементами других локально-оптимальных решений.Поиск в окрестностях, образованных добавлением в известное промежуточное решение, представленное парой множеств <*D’,W’*>, элементов другого решения <*D’’,W’’*> с последующим удалением «лишних» решений жадной агломеративной эвристической процедурой выполняется Алгоритмами 4, 5, 6. Эти алгоритмы осуществляют поиск в некоторых окрестностях решения <*D’,W’*>, а второе решение <*D’’,W’’*> задает параметры этой особой окрестности. Так, Алгоритм 4 осуществляет поиск в окрестности

$$S(<D',W'>)=\{\text{greedy}(<D',W'>\cup\{<D_i'',W_i''>\}),i=1,|D''|\}.$$

Здесь *greedy()* – результат применения Алгоритма 3. Окрестности в Алгоритмах 5 и 6 шире. Такая стратегия, фактически осуществляя расширенный локальный поиск в окрестностях, выбор которых рандомизирован, используется в качестве цикла А метода жадных эвристик (традиционно - цикл глобального поиска).

Алгоритм 7. Поиск в чередующихся окрестностях для разделения смеси распределений (расширенный локальный поиск).

1. Запустить ЕМ-алгоритм (локальный поиск) из случайного начального решения, получить решение <*D,W*>.
2. Установить $s=s_{start}$ (номер типа окрестностей)
3. Установить $i=0$, $j=0$; (количество безрезультатных итераций в конкретной окрестности и в целом по алгоритму).
4. Запустить ЕМ-алгоритм (локальный поиск) из случайного начального решения, получить решение <*D’,W’*>.

5. В зависимости от значения s (возможны значения 1, 2 или 3), запустить Алгоритм 4, 6 или 5 с начальными решениями <*D,W*> и <*D',W'*>.

6. Если результат по значению целевой функции лучше, чем <*D,W*>, то заменить <*D,W*> этим новым результатом, присвоить i=0, j=0, перейти к Шагу 5.

7. Присвоить $i=i+1$;

8. Если $i<i_{max}$, то перейти к Шагу 4.

9. Присвоить $i=0$, $j=j+1$, $s=s+1$; если s>3, то присвоить s=1;

10. Если $j>j_{max}$, или выполняются другие условия останова (например, максимальное время работы), то ОСТАНОВ. Иначе перейти к Шагу 5.

Мы использовали значение $i_{max}=2k$, $j_{max}=2$.

Также важным является параметр s_{start}, задающий номер окрестности, с которой начинается поиск. В зависимости от этого значения алгоритмы обозначены ниже соответственно VNS1, VNS2, VNS3. Эксперименты показывают высокую конкурентоспособность различных версий Алгоритма 7 для определенных классов задач.

Для тестирования новых алгоритмов использовались как результаты тестирования ЭРИ, полученные ОАО «ИТЦ НПО-ПМ» (эти данные нормированы специальным образом – по границам дрейфа), так и классические наборы данных из репозитория UCI. Результаты в таблице 1. Использовалась вычислительная система с процессором XeonX5650 2.67 GHz, 12GbRAM.

Результаты показывают, что новые алгоритмы не имеют преимущества перед мультистартом классического EM-алгоритма при очень малом числе кластеров (до 5 кластеров). С ростом числа кластеров и объема выборки сравнительная эффективность повышается, и для многих больших задач новые алгоритмы имеют безусловное преимущество. Также новые алгоритмы оказались неэффективны для разбиения наборов булевых данных (см. набор данных Chess).

Часть второй главы посвящена разработке генетических алгоритмов для решения нашей задачи. При этом в их основу ложатся те же жадные процедуры.

Эволюционные алгоритмы, включая генетические, показывают высокую эффективность при решении задач автоматической группировки на основе модели k-средних и аналогичных. Сложности кодирования решений, традиционно представляемых в классических генетических алгоритмах L-битными строками, в алгоритмах метода жадных эвристик решены применением так называемого генетического алгоритма с вещественным алфавитом, в котором «особи» - промежуточные решения – представлены непосредственно множествами точек в пространстве R^d.

Аналогичный подход с кодированием промежуточных решений в виде множеств векторов вещественных чисел применялся и при построении генетического алгоритма для решения задач разделения смесей распределений. В нашем алгоритме промежуточные решения представлены парами множеств $D_l = \left\{N(\mu_{l,i}, \sigma_{l,i}{}^2 I_n), l=\overline{1,k}\right\}, l=\overline{1,N_{POP}}$ и $W_l = \left\{\alpha_{l,i}, i=\overline{1,k}\right\}$, где N_{POP} – размер популяции алгоритма, т.е. количество «особей» - промежуточных решений, которым он оперирует. Могут использоваться и другие виды распределений.

Таблица 1–Сравнительные результатыизвестных алгоритмов и новых с чередующимися окрестностями

	MULT	**CEM**	**SEM**	**VNS1**	**VNS2**	**VNS3**
Chess (UCI), N=3197, d=50, булевы, сфер.,2 мин.,10 кластеров						
F(x)	-30525,0∗	-30564,1	-30560,7	-30554,7	-30539,9	-30580,6↓
σ	0,0	32,3	46,4	28,7	25,4	0,0
ИС 140УД17АВК, N=51, d=46, сфер.,5 сек., 5 кластеров						
F(x)	3598,1↓			3673,7*	3543,9	3591,4
σ	32,2			44,0	144,7	139,9
Ionosphere, N=351, d=35, нормирован.,сфер., 2 мин.,10 кластеров						
F(x)	-871,4	-893,4↓	-880,0	-847,5*	-849,4	-878,2
σ	15,8	7,9	15,2	24,0	28,1	41,9
Mopsi (UCI), N=6014, d=2, сфер.,10 мин., 20 кластеров						
F(x)	39268,7	48424,2	36272,2↓	50291,1	50311,6	50370,8*
σ	9967,6	237,3	9619,6	122,9	104,7	51,0
Europe, (UCI), N=169308, d=2, некорр.,15 мин., 30 кластеров						
F(x)	-3653148,6		-3654493↓	-3648660,1	-3644645,2	-3639340∗
σ	2634,7		4348,6	1383,5	1840,9	1729,9
BIRCH-3, N=100 000, d=2, сфер.,15 мин., 10 кластеров						
F(x)	-2704718↓	-2693183,4	-2694786,1	-2701230,1	-2700481,8	-2688604*
σ	4738,8	3890,8	2065,6	5567,4	5589,4	0,0
KDDCUP04Bio (UCI), N=145751, d=74, нормир.,сфер., 60 мин., 30 кластеров						
F(x)	-12513229	-12512717	-12514336↓	-12511968∗	-12511984	-12512654
σ	1255,5	343,0	683,6	105,3	278,0	139,0
Missamerica (UCI), N=6480, d=16, сфер., 60 мин.,10 кластеров						
F(x)	-267008↓	-266378,7	-266981,3	-266553,3	-266616,9	-266351∗
σ	141,5	0,6	52,5	273,6	400,6	0,0
Missamerica (UCI), N=6480, d=16, сфер., 60 мин.,30 кластеров						
F(x)	-261971,1	-262265↓	-261839,6	-261741,9	-261737,9	-261737*
σ	99,6	51,6	26,9	10,0	8,7	1,5

Примечания: ∗-лучший результат; ↓ - худший результат.

Алгоритм 8. Генетический алгоритм с жадной эвристикой (приведено общее описание трех вариантов, названных GA-FULL, GA-ONE, GA-RAND). Дано: Начальный размер популяции N_{POP}.

1. Сгенерировать случайным образом N_{POP} начальных решений, представленных множествами распределений $D_l = \left\{N(\mu_{l,i}, {\sigma_{l,i}}^2 I_n), l = \overline{1,k}\right\}, l = \overline{1, N_{POP}}$ и соответствующими множествами весовых коэффициентов $W_l = \left\{\alpha_{l,i} = 1/k, i = \overline{1,k}\right\}$. Начальные значения среднеквадратичных отклонений устанавливаются равными для всех кластеров и вычисляются для всей выборки: ${\sigma_i}^2 = \frac{1}{d}\sum_{x \in S}\left\|x - \bar{x}\right\|^2$. Значения $\mu_{l,i}$ устанавливаются равными координатам случайно выбранных векторов данных.

Для каждого из начальных решений запускается Алгоритм 1, полученные значения целевой функции сохраняются в переменных $f_1,...,f_{N_{POP}}$. ПрисвоитьN_{iter}=0;

2. Проверка условий останова (например, максимальное время работы). Если условия достигнуты, возвращается решение, которому соответствует наилучшее (наибольшее) значение целевой функции $f_1,...,f_{N_{POP}}$.

3. Присвоить $N_{iter}=N_{iter}+1$; $N_{POP} = \max\{N_{POP}; \lceil\sqrt{1+N_{iter}}\rceil + 2\}$; если N_{POP} изменилось, то инициализировать новое решение (см. Шаг 1).

4. Выбрать случайным образом два индекса $k_1,k_2 \in \{\overline{1,N}\}, k_1 \neq k_2$. Для пары решений, представленных множествами D_{k_1}, D_{k_2} и W_{k_1}, W_{k_2}, выполнить Алгоритм 5 (в варианте алгоритма GA-FULL – генетический алгоритм с жадной эвристикой с полным объединением), Алгоритм 4 (в варианте алгоритма GA-ONE – генетический алгоритм с жадной эвристикой с частичным объединением), либо выбрать один из двух этих алгоритмов случайным образом (вариант алгоритма GA-RAND).

5. Выбрать индекс $k_3 \in \{\overline{1,N_{POP}}\}$. Используем простое турнирноезамещение: случайным образом выбираем $k_4,k_5 \in \{\overline{1,N_{POP}}\}$, если $f_{k_4} < f_{k_5}$ то $k_3=k_4$, иначе $k_3=k_5$

6. Заменяем множества D_{k_3}, W_{k_3} и соответствующее значение целевой функции f_{k_3} новыми значениями, полученными на Шаге 4. Перейти к Шагу 2.

Результаты даны в таблице 2. Для большинства наборов данных было выполнено по 30 попыток запуска каждого из алгоритмов. Фиксировались только лучшие результаты, достигнутые в каждой попытке, затем эти результаты были усреднены. Результаты работы EM-алгоритма в режиме мультистарта и его модификаций обозначены: EM, CEM, SEM. Как и в случае алгоритмов с чередующимися окрестностями, новые генетические алгоритмы неэффективны для малых задач и задач с небольшим числом кластеров. Для большинства больших задач, в особенности для задач с векторами данных большой размерности (десятки и сотни измерений) именно генетические алгоритмы показывают наилучшие результаты при достаточном времени счета. При уменьшении времени счета преимущество за алгоритмами с чередующимися окрестностями.

В зависимости от требований к точности и стабильности результата, допустимого времени работы алгоритма и особенностей конкретного класса решаемых задач при построении систем автоматической группировки предлагается составлять комбинации алгоритмов путем выбора из следующих множеств алгоритмов:

А. Мультистарт базовой жадной эвристики (Алгоритм 3) или VNS (Алгоритм 7) или ГА (Алгоритм 8);

Б. Одна из модификаций жадных эвристик (Алгоритм 4 или 5 или 6);

В. EM или CEM или SEM и т.д.

Глава 3 посвящена разработке модели разделения партии промышленной продукции. Результаты неразрушающего тестирования промышленной продукции, к которой предъявляются повышенные требования качества, как правило, представлены векторами числовых данных – результатов замеров тех или иных параметров. Кроме того, могут проводиться выборочные разрушающие испытания. Результаты разрушающих испытаний могут быть распространены на всю партию продукции лишь при условии, что вся партия является однородной, произведенной в одних и тех же условиях и из единой партии сырья. Если такой убежденности нет, результаты

неразрушающих тестов могут быть использованы для разделения сборной партии изделий на однородные партии. Так, результаты тестирования ЭРИ наиболее высокого уровня качества, проводимых в специализированных тестовых центрах в ходе входного контроля, дополнительных отбраковочных испытаний и дополнительного неразрушающего контроля, представлены векторами числовых данных большой размерности: число измерений, проводимых в ходе перечисленных видов испытаний, достигает нескольких сотен.

Таблица 2–Сравнение результатов известных алгоритмов и новых генетических

	EM	**CEM**	**SEM**	**GA-FULL**	**GA-ONE**	**GA-RAND**
Chess (UCI), N=3197, d=50, булевы, сфер.,3 мин., 10 кластеров						
F(x)	-30525,0*	-30564,1	-30560,7	-30581,1↓	-30532,4	-30554,7
σ	0,0	32,3	46,4	1,8	19,6	28,7
ИС 140УД17АВК, N=51, d=46, сфер.,0,5 мин., 2 кластера						
F(x)	3790,1*			3665,6↓	3697,8	3707,7
σ	104,4			0,0	83,4	88,0
ИС 140УД25АС1ВК, N=56, d=43, некорр.,2 мин., 10 кластеров						
F(x)	7541,0	853,6 ↓	928,9	7692,4*	7291,7	7551,5
σ	457,9	7,9	105,5	1335,8	473,7	438,5
Ionosphere, N=351, d=35, нормирован.,сфер., 1 мин.,10 кластеров						
F(x)	-871,4	-893,4	-880,0	-960,3↓	-824,8*	-832,8
σ	15,8	7,9	15,2	23,5	4,5	14,8
Mopsi (UCI), N=6014, d=2, сфер.,15 мин., 20 кластеров						
F(x)	39268,7	48424,2	36272,2↓	50443,4	49288,9	50499,9*
σ	9967,6	237,3	9619,6	115,3	390,5	60,9
Europe, (UCI), N=169308, d=2, некорр.,60 мин., 30 кластеров						
F(x)	-3653148,6	-	-3654493↓	-3643277*	-3651917,0	-3648896,2
σ	2634,7	-	4348,6	2122,9	3465,7	5151,5
BIRCH-3 (UCI), N=100 000, d=2, сфер.,60 мин., 100 кластеров						
F(x)	-2567483,0	-2603150,9	-2728547↓	-2553037,9	-2348371*	
σ	4351,1	5170,3	3869,7	8014,0	588278,1	
KDDCUP04Bio (UCI), N=145751, d=74, нормир.,сфер., 90 мин.,30 кластеров						
F(x)	-12513229	-12512717	-12514336↓	-12512517*	-12512817	-12512811
σ	1255,5	343,0	683,6	159,1	410,9	639,1
MissAmerica (UCI), N=6480, d=16, сфер., 40 мин., 30 кластеров						
F(x)	-261971,1	-262265↓	-261839,6	-261736*	-261893,5	-261763,1
σ	99,6	51,6	26,9	1,6	71,7	22,3

Примечания: *-лучший результат; ↓ - худший результат.

Имея результаты испытаний заведомо однородной партии (рис. 1), изготовленной из однородной партии сырья, можно оценить характер вероятностного распределения каждого из измерений (параметров) ЭРИ в партии. Некоторые результаты по оценке характера распределений достигнуты в работах В.В.Федосова и В.И.Орлова. Также показано, что значения дрейфа параметров, полученные исходя из измерений до и после электротермотренировки, имеют вид, близкий к нормальному.

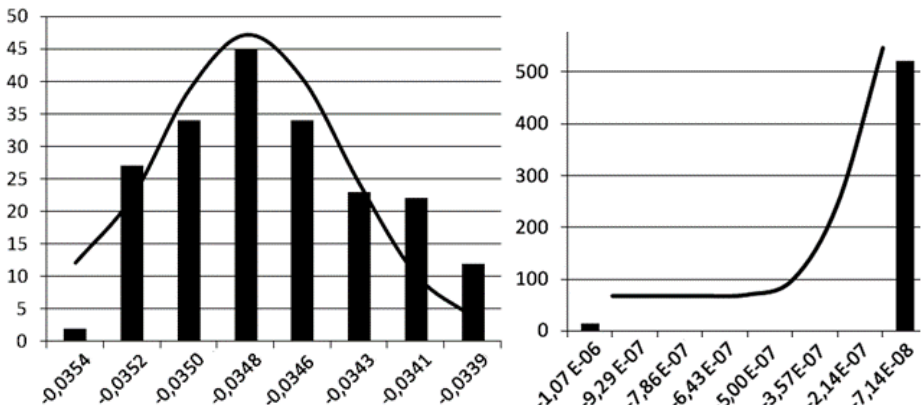

Рисунок 1 – Частотная гистограмма параметров. Слева - параметр Т2 ИС 1526ТЛ1, партия №1 (199шт.), справа – параметр Т13, партия №2 (535 шт.)

Анализ частотных гистограмм показывает, что характер распределений параметров в различных партиях одинаков, такие параметры, как среднеквадратичное отклонение – соизмеримы. Все измеренные в ходе неразрушающих тестов параметры можно разделить на 3 группы:

А) Параметры, для которых нормальных характер распределения четко прослеживается (рис.1, слева).

Б) Параметры, для которых нормальный характер распределения прослеживается на частотной гистограмме, но точность измерительных приборов такова, что числовой ряд возможных значений данного измерения распадается на небольшое по мощности конечное множество значений, которые может показать измерительный прибор. В этом случае частотная гистограмма содержит существенные пропуски, и оценка правдоподобия нормальности распределения критерием Пирсона дает отрицательный ответ.

В) Параметры, для которых частотная гистограмма имеет вид, не соответствующий нормальному распределению (рис.1, справа).

Критерий эксцесса $C_{ex} = \frac{\sum_{i=1}^{N}(x_i - \frac{1}{N}\sum_{j=1}^{N} x_j)^4}{(\sum_{i=1}^{N}(x_i - \frac{1}{N}\sum_{j=1}^{N} x_j)^2)^2} - 3 = \frac{\sum_{i=1}^{N}(x_i - \overline{x})^4}{\sigma^4} - 3$ позволяет отделить параметры группы В от остальных. Здесь $x_1 \ldots x_N$ – значения измерений некоторого параметра изделий партии из N штук, $\overline{x}$ – среднее значение этого параметра, σ – среднеквадратичное отклонение. Для таких параметров данный критерий принимает высокие значения, более 10. Для нормально распределенной случайной величины данный критерий имен нулевое математическое ожидание.

Отметим, что для различных изделий параметры группы А составляют весьма значительную часть всего множества параметров: от 40% до 70%.

Модель сборной партии ЭРИ в виде смеси сферических или некоррелированных гауссовых распределений позволяет разделять составные части сборных партий, при этом процент ошибок достаточно велик. Модель разделения партий ЭРИ на основе модели k-средних также достаточно хорошо исследована в работах В.И.Орлова, Л.А.Казаковцева, А.Н.Антамошкина, при этом показано, что нормировка исходных данных с использованием установленных допустимых границ дрейфа параметров позволяет снизить процент ошибок при разделении сборной партии. Именно для параметров группы А и параметров группы Б в основном установлены допустимые границы дрейфа параметров в ходе электротермотренировки. Отделив применением критерия эксцесса параметры, нормальность распределения которых не подтверждается, можно сформировать

выборку, параметры которой с высокой вероятностью имеют нормальное распределение. К такой выборке можно применять модель разделения сборной партии изделий на основе разделения смеси гауссовых распределений.

Для проверки адекватности модели разделения сборной партии ЭРИ в качестве тестового примера использовались две сборные партии: ИС 1526ИЕ10 из 870 штук, составленная из четырех партий и сборная партия ИС 140УД25АВК из 53 штук.Отсев параметров, имеющих распределение, отличное от нормального, позволяет повысить адекватность модели разделения сборных партий электронных устройств на основе смеси гауссовых распределений (таблица 3). Для разделения смеси распределений использовались новые генетические алгоритмы.

Таблица 3– Сравнительные результаты разбиения сборной партии ИС 1526ИЕ10, сборная партия №2 (870 шт.).

Модель	Партия №1		Партия №2		Партия №3		Партия №4		% ошибок
	Кол-во согл.мо-дели	Ош.отнесенные к др. партиям	Кол-во согласно модели	Ош.отнесенные к др. партиям	Кол-во согласно модели	Ош.отнесенные к др. партиям	Кол-во согласно модели	Ош.отнесенные к др. партиям	
Фактически	146	-	229	-	424	-	71	-	-
Некор. гаус. распр. по всем измерениям	244	25	196	41	358	78	72	1	16,7%
Новая –некор. гаус. распр. с отбраковкой измерений критерием эксцесса	125	26	221	8	414	10	110	0	5,05%
Сферич.гауссовы распр.	251	0	189	40	359	76	71	0	13,3%

Таким образом, новая модель позволяет вдвое снизить процент ошибок разделения сборной партии, а новые алгоритмы позволяют применять данную модель, получая максимально точный (по значению функции правдоподобия) и воспроизводимый при многократных перезапусках алгоритма (а следовательно – проверяемый) результат.

ЗАКЛЮЧЕНИЕ

В диссертации разработаны новые алгоритмы для решения задач автоматической группировки объектов на основе разделения смеси вероятностных распределений, позволяющие успешно решать широкий круг практических задач с повышенной точностью (по целевой функции правдоподобия) за ограниченное время, позволяющее использовать новые алгоритмы в составе интерактивных систем автоматической группировки объектов. Сравнительная эффективность новых алгоритмов доказана экспериментально на различных наборах данных.

Цель диссертации достигнута путем решения поставленных задач, а именно:

1. Анализ известных методов автоматической группировки, основанных на модели разделения смеси распределений, выявил, что, несмотря на имеющийся широкий набор известных алгоритмов решения задачи, все они являются методами локального поиска, не гарантирующими не только точного решения, но и решения, которое было бы трудно улучшить за ограниченное время. Анализ методов, позволяющих повысить получаемые значения функции правдоподобия и

стабильность этих значений для задач, основанных на моделях теории размещения, выявил общность постановок задач и общность структуры наиболее типичных алгоритмов, что дало обоснованную надежду на применение аналогичных подходов к повышению точности получаемых результатов, в частности – подхода с применением идей метода жадных эвристик для систем автоматической группировки объектов;

2. Впервые предложены эффективные алгоритмы решения задач автоматической группировки объектов на основе разделения смеси распределений в пространствах большой размерности с применением жадных агломеративных эвристических процедур и идей метода поиска с чередующимися окрестностями. Показано, что новые алгоритмы, в которых в качестве стратегии расширенного локального поиска применен оригинальный подход с поиском в чередующихся окрестностях, позволяют получать более точный по значению целевой функции результат по сравнению с известными методами, для задач с наборами данных большой размерности (сотни измерений) за ограниченное время, приемлемое для работы в интерактивном режиме.

3. Построены новые эффективные генетические алгоритмы метода жадных эвристик для решения больших задач автоматической группировки объектов на основе разделения смеси распределений. Показано, что новые генетические алгоритмы позволяют получать наиболее точный по значению целевой функции результат по сравнению с известными методами за ограниченное время для больших задач (до сотен тысяч векторов данных) с наборами данных большой размерности (сотни измерений).

4. Предложен подход к составлению эффективных комбинаций алгоритмов для систем автоматической группировки объектов на основе разделения смеси распределений, являющийся развитием метода жадных эвристик.

5. Построена новая модель разделения сборных партий промышленной продукции на однородные партии на основе модели разделения смеси распределений. На примере задачи выявления однородных производственных партий ЭРИ из сборной партии по данным неразрушающих тестовых испытаний показано, что новая модель позволяет снизить процент ошибок при выявлении однородных партий.

ПУБЛИКАЦИИ ПО ТЕМЕ ДИССЕРТАЦИИ

В издании, входящем в систему цитирования Scopus:

1. Stashkov D.V. Fuzzy clustering of EEE components for space industry / Orlov V.I., Stashkov D.V., Kazakovtsev L.A., Stupina A.A. // IOP Conf. Series: Materials Science and Engineering.–2016.–Vol.155.–article ID 012026, 10 P. DOI: 10.1088/1757-899X/155/1/012026.

В изданиях из перечня ВАК:

2. Сташков Д.В. Выявление однородных партий изделий космической радиоэлектроники на основе разделения смеси сферических гауссовых распределений В. И. Орлов, Сташков Д.В., Л. А. Казаковцев, И. Р. Насыров, А. Н. Антамошкин// Вестник СибГАУ. – 2017.Т.18. – вып. 1. – С. 69-77.

3. Сташков Д.В. Задача формирования отказоустойчивого программного комплекса управления промышленным объектом /Сташков Д.В., Насыров И.Р., Казаковцев Л.А. // Фундаментальные исследования.— 2015.— вып. 5.1 — С.149-153.

4. Сташков Д.В. Алгоритмы поиска с чередующимися окрестностями для задачи разделения смеси распределений// Системы управления и информационные технологии, №1(67), 2017. – С. 18-24.

5. Сташков Д.В. Применение ЕМ-алгоритма со сферическим гауссовым распределением к задаче классификации промышленной продукции/ Сташков Д.В., Орлов В.И., Насыров И.Р., Казаковцев Л.А. // Экономика и менеджмент систем управления. — 2017. — №1.1(23) С. 185-193.

6. Сташков Д.В. Моделирование сборной партии электрорадиоизделий в виде смеси вероятностных распределений/ Сташков Д.В., Насыров И.Р., Попов А.М., Орлов В.И. // Экономика и менеджмент систем управления. — 2017. — №2.2(24) С. 282-289.

В рецензируемой монографии:

7. Сташков Д.В. Эволюционные алгоритмы с гетерогенной популяцией для задач кластеризации и размещения: монография / Л.А.Казаковцев, М.Н.Гудыма, Д.В.Сташков, А.А. Ступина, Н.Н. Джиоева; Сибирский федеральный ун-т. – Москва: Издательство НИЦ «Актуальность.РФ». – 196 с.

В других изданиях:

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

9. Сташков Д.В. Алгоритм разделения смеси сферических гауссовых распределений повышенной точности/ Д.В.Сташков, Л.А. Казаковцев, В.И. Орлов, А.Н. Антамошкин // Системный анализ и информационные технологии, г. Светлогорск, 13-18 июня 2017, С.438-445.

10. Сташков Д.В, Насыров И.Р. Мультиверсионное программирование в автоматических системах управления технологическими процессами // Приоритетные научные направления: от теории к практике.— 2015. — вып. 16.— С.75-78.

11. Сташков Д.В., Насыров И.Р. ЕМ-алгоритм для разделения смеси сферических гауссовых распределений // Информационные технологии моделирования и управления.—2016. Т. 102.— № 6.—С. 464-471.

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

13. Stashkov D.V. Fuzzy clustering algorithm for multidimension data // Молодежь. Общество. Современная наука, техника и инновации.–2017.– № 16 С. 354-356.

14. Stashkov D., Kazakovtsev L. Genetic fuzzy clustering algorithm with greedy crossing-over procedure // Methods of Optimization and Their Applications. Irkutsk, 31 July – 6 August 2017, P.124.

Свидетельство о государственной регистрации программ для ЭВМ:

1. Сташков Д.В. Система интеллектуального анализа данных на основе алгоритмов разделения смеси гауссовых распределений повышенной точности/ Д.В.Сташков, Л.А. Казаковцев, В.И.Орлов, В.Л.Казаковцев, В.В.Проценко.– М.: РОСПАТЕНТ.–2017.–Свидетельство о государственной регистрации программы для ЭВМ № 201766152.