Федеральное государственное бюджетное образовательное учреждение
высшего образования «Сибирский государственный университет
науки и технологий имени академика М.Ф. Решетнева»

На правах рукописи

Сташков Дмитрий Викторович

СИСТЕМЫ АВТОМАТИЧЕСКОЙ ГРУППИРОВКИ ОБЪЕКТОВ НА ОСНОВЕ РАЗДЕЛЕНИЯ СМЕСИ РАСПРЕДЕЛЕНИЙ

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

ДИССЕРТАЦИЯ
на соискание ученой степени кандидата технических наук

Научный руководитель
доктор технических наук, доцент
Казаковцев Л.А.

Красноярск – 2017

Оглавление

ВВЕДЕНИЕ

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.

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.

В случае наиболее распространенных моделей, основанных на разделении смеси многомерных нормальных (гауссовых) распределений и их частных случаев, таких как некоррелированные и сферические гауссовы распределения, а также некоторые другие (например, распределения Лапласа), функция правдоподобия является многоэкстремальной функцией в пространстве возможных параметров 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). Алгоритмы популярны, являются де-факто стандартом при построении автоматизированных систем, реализующих группировку объектов на основе разделения смеси распределений. Несмотря на это, алгоритм имеет ряд существенных недостатков. В частности, он крайне неустойчив по отношению к исходным данным. Кроме того, алгоритм требует задания предполагаемого числа распределений в смеси. Более того, алгоритм и его модификации требуют задания начального решения, являясь (не в строгом смысле) алгоритмами локального поиска. Алгоритм неустойчив к выбору начального решения. Лишь частично данная проблема решается применением SEM-алгоритма, который выполняет случайное «встряхивание» выборки на каждой итерации.

Как показывают вычислительные эксперименты, стабильность результатов при многократных запусках SEM-алгоритма выше, чем у EM-алгоритма, в то же время результат во многих случаях далек от истинного оптимума функции правдоподобия. 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 EM-алгоритм.

Систематизация постановок задач кластеризации и алгоритмов их решения, включая EM-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, гидроакустике и во многих других прикладных областях.

Процедура, схожая с EM-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 «EM-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 EM-алгоритма с повышенной стабильностью результата с использованием медианных

модификаций EM-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 EM-подход и метод Ньютона. Ч.Лиу, Д.Б.Рубин и др. предложен метод, позволяющий улучшить скорость работы EM-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), задач кластеризации и размещения (J.Brimberg, N.Mladenovicetal., 1996) и других. Согласно «гипотезе о большой долине», они 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 основе различных 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 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;

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 наилучшие значения 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 нормировки 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 статьях и 1 рецензируемой монографии, среди которых 5 работ в ведущих российских рецензируемых изданиях, рекомендуемых в действующем перечне ВАК, 1 - в международных изданиях, проиндексированных в системе цитирования Scopus.

ГЛАВА 1. ОБЗОР МЕТОДОВ РЕШЕНИЯ ЗАДАЧ АВТОМАТИЧЕСКОЙ ГРУППИРОВКИ ОБЪЕКТОВ И РАЗДЕЛЕНИЯ СМЕСИ РАСПРЕДЕЛЕНИЙ

1.1 Общая постановка задач автоматической группировки объектов

Целью автоматической группировки данных, т.е. кластерного анализа, ставится обнаружение естественной группировки ряда образцов, пунктов или объектов. Решение задачи автоматической группировки сводится к разработке алгоритма и/или автоматизированной системы, способных обнаруживать эти естественные группировки в не маркированных предварительно данных.

В анализе данных выделяют два основных типа методов [200]: исследовательский или описательный, в котором исследователь без использования заранее определенных моделей и гипотез пытается выделить общие характерные свойства многомерных данных, и логически выведенный- подтверждающий тип методов, для которого необходимо подтверждение адекватности модели или справедливости предположений на имеющихся данных.

В процессе развития направления анализа данных появилось много статистических методов: линейная регрессия, дискриминантный анализ, дисперсионный анализ, многомерное шкалирование, анализ корреляции, факторный анализ, кластерный анализ [196]. Для задач распознавания образов целью анализа данных ставится построение прогнозной модели: предсказание поведения данных по их части. Задачи в такой постановке также называются обучением. Задачи обучения делятся на два класса: обучение с учителем, т.е. классификация (имеется обучающая выборка с данными, для которых заранее определена принадлежность к тому или иному классу) и обучение *без учителя* (объединение в кластеры при полном отсутствии априорной информации о принадлежности даже части данных к конкретной группе; в практических задачах зачастую даже неизвестно число таких кластеров) [120].

Задачи автоматической группировки можно отнести к любому из этих классов: в обоих случаях должна решаться задача разбиения множества объектов

на однородные группы со схожими характеристиками. Следует отметить, что объединение в кластеры является задачей более сложной. Развивается гибридный подход - полуконтролируемое обучение [188], при котором вхождение объектов в группу (маркировка) известно только для части обучающей выборки. Разделение оставшихся элементов по группам выполняется на после анализа близости элементов к маркированным объектам. Полуконтролируемая кластеризация заменяет четкое разбиение объектов на классы (маркировка объектов) на нахождение попарных ограничений. Такие попарные ограничения являются более "слабым" способом закодировать предварительные знания [160] и сводятся к требованиям о том, что два элемента (объекта данных) обязаны быть объединены в одну группу (кластер), либо наоборот не могут входить в одну группу (кластер).

Задача *четкой* кластеризации можно сформулировать следующим образом: используя информацию о N объектах, выделить в них k групп (т.е. выполнить разбиение N объектов на k непересекающихся подмножеств), основываясь на некой мере подобия таким образом, чтобы объекты, принадлежащие одной и той же группе, были подобны друг другу, то есть обладали схожими параметрами или характеристиками, а объекты, принадлежащие различным группам, были не схожи друг с другом.

Задача нечеткой кластеризации состоит в отнесении каждого из Nобъектов к каждой из k групп с некоторой условной вероятностью.

Формулирование задачи разбиения объектов на группы в таком виде порождает как минимум два вопроса: как определить количество групп объектов k, и какую меру подобия использовать.

Задачи автоматической группировки в литературе упоминаются как таксономия, Q-анализ, типология, оптимальное разбиение объектов на группы и т.п. Следует заметить, что данные понятия не являются тождественными.

Такого рода задачи возникают в любой дисциплине, связанной с анализом данных многомерной структуры.

Примерами задач автоматической группировки являются: сегментация изображения (в компьютерном зрении) [145, 7, 190, 12], группировка документов

(организация быстрого доступа и поиска, эффективного использования памяти при хранении) [143, 8, 9, 97], группировка клиентов по географической близости для оптимальной организации обслуживания, задачи распознавания печатного [65] и рукописного [113] текста, управления ресурсами и планирования [142], биологические задачи [91, 5].

Автоматическая группировка широко используется при естественной классификации (например, природных объектов или производственных партий электрорадиоизделий, рассматриваемых в Главе 3), для структурирования данных, выделении аномалий в данных (например, некачественной продукции различных производственных процессов), для сжатия данных (выполняется замена близких по характеристикам или одинаковых объектов данных на единственный объект, который можно считать их обобщенным представлением.)

Различные подходы подразумевают использование всевозможных мер сходства, различных целевых функций (задачи минимизации суммарного расстояния между объектами, до центров кластеров; минимизация максимальных расстояний в кластере и т.д.), различных эвристических процедур и схем для организации локального и глобального поиска. Применение данных подходов обусловлено невыпуклостью целевых функций большинства таких задач. В Главе 2 данной работы приводится обзор и обобщение стратегий глобального и локального поиска, применяемых совместно с жадными агломеративными эвристическими алгоритмами.

В подходах, основанных на оценке и анализе плотности вероятности, группы объектов определяются как области высокой плотности в пространстве признаков (характеристик) на фоне областей низкой плотности. Использование определения плотности и поиска связанных областей высокой плотности в пространстве признаков дало основу для класса алгоритмов. При этом используются различные определения связности для класса плотностных алгоритмов. Алгоритмы, подобные алгоритму Джарвиса-Патрика, определяют подобие пары точек как число их общих соседей, разделенными этой парой (соседи являются точками внутри круга заданного радиуса вокруг точки), либо определяют плотность, применяя метод окна

Парзена [124, 39]. Гистограмма и полиграмма низших порядков являются простейшими оценками плотностей [58, 66]. В работах Е. Парзена [176, 177], В.А. Епанечникова [21], А.В.Медведева [46] предложены методы непараметрического оценивания плотности. В работах этих и других авторов были введены классы оценок для обобщения гистограммы. Один из таких классов был предложен Розенблаттом и Парзеном и назван классом "ядерных оценок". Исследование скорости сходимости отклонений, обеспечения несмещенности таких оценок, изучение влияние формы ядра на качество приближения оценки к функции плотности и рассмотрение вопросов определения параметров алгоритмов восстановления плотности — параметра размытости (сглаживания) непараметрических оценок плотности, например — с гауссовым ядром приводятся в [46, 50, 184, 25]. Для задач с большим объемом входных данных предложены [123] принципы построения систем автоматической группировки, позволяющие обрабатывать данные за один проход. Следует заметить, что алгоритмы, основанные на методах восстановления плотности, имеют зависимость от выбранного масштаба измерения расстояния, от количества точек в окрестности друг друга для определения такого скопления точек как группы, а также от максимальной удаленности между точками в группе. Определение этих параметров является обособленной и сложной задачей. От качества ее решения зависит точность метода, а также достоверность модели автоматической группировки. Реализация подобных методов сталкивается с существенной проблемой - *«проклятием размерности»* [96]. Эта проблема состоит в том, что необходимый объем данных этих алгоритмов, при котором они эффективно восстанавливают плотность, экспоненциально зависит от размерности пространства. Для случаев, когда имеется относительно небольшой объем данных, а число учитываемых признаков велико, то все объекты оказываются примерно на одинаковом расстоянии друг от друга. В связи с этим фактом использование подобных подходов для задач с многомерным данным практически неприменимо.

Предложено множество подходов на основе вероятностных моделей плотности [168, 98, 38], например, ненаправленная графическая модель [205] и

модель размещения Патинко [162]. Такие методы имеют те же ограничения на размерность данных, выбор параметров, они требовательны к вычислительным ресурсам. Но, несмотря на это, такие методы (в особенности, основанные на непараметрических моделях плотности [38]) имеют популярность благодаря своей способности выделять кластеры произвольной формы. В случаях, когда имеется очень высокая размерность пространства признаков, сравнимая с количеством группируемых объектов, данные (объекты, точки) в таком пространстве удалены друг от друга, как правило, весьма значительно. Из-за этого построение универсальной процедуры выбора параметров этих алгоритмов для достижения эффективной группировки практически невозможно. Для задач с небольшим объемом входных данных разработаны непараметрические алгоритмы, использующиеся в комбинации с другими популярными методами, например, с методом k-средних. В [172] рассматриваются примеры с 2 кластерами.

Задачи автоматической группировки и размещения традиционно формулируются российскими (ранее советскими) исследователями в непрерывном пространстве и на сетях задачами целочисленного линейного программирования [64]. Эти задачи NP-трудны, но несмотря на это, для них разработан большой арсенал в основном точных эффективных методов решения [10, 17, 18, 15, 3, 1, 13, 23]. Представлены методы решения и для двухуровневых задач с оценкой значения целевой функции в процессе решения вложенной оптимизационной задачи. Из-за резкого увеличения объемов данных, обрабатываемых в автоматизированных системах, для алгоритмов, ранее эффективных при решении задач, NP-трудность решаемых задач автоматической группировки становится проблемой. NP-трудными задачами считается класс задач, которые «как минимум так же трудны, как самая трудная из задач класса «NP». То есть, NP-трудными называют задачи, к которым можно свести любую задачу из класса NP за время, полиномиально зависящее от объема входных данных. К классу NP относятся задачи, решения которых могут быть проверены на недетерминированной машине Тьюринга за полиномиальное время. *P*-медианная задача на графе и подобные ей задачи, кроме NP-трудности, обладают тем свойством, что за полиномиальное время в общем

случае невозможно получить приближенное решение с постоянной оценкой точности [90].

При иерархическом методе группировки выполняется объединение объектов в непрерывном пространстве характеристик, при этом строится древовидная модель взаимоотношений объектов. Такой подход в какой-то мере проявляет сходство иерархической группировки с методами автоматической группировки на графах [128]. Данные методы используются при решении задач группировки при достаточно больших объемах данных. Алгоритмы иерархического метода автоматической группировки выполняют рекурсивное агломеративное выделение вложенных групп (сначала каждый объект рассматривается в качестве отдельной группы, затем происходит последовательное объединение этих групп попарно, таким образом формируется иерархия этих групп), или применяется аналитический (нисходящий, диссоциативный) способ, начинающийся со всех объектов данных в одной группе и рекурсивно делящий каждую группу на меньшие группы. В алгоритмах, подобных методу k-средних, одновременно определяются все группы, не используя иерархическую структуру. Данные для иерархического метода группировки представляют собой матрицу подобия размера N x N, где N - число объектов группировки. При использовании методов, подобных k-средних, объекты должны располагаться в d-мерном пространстве характеристик. Можно выделить разновидности алгоритмов, таких как k-медоид, способных работать и с матрицей подобия. Получение матрицы подобия по координатам объектов в пространстве характеристик не вызывает сложностей, но преобразование в обратную сторону сопряжено с применением сложных методов, например, многомерного шкалирования [144] (данные методы сами могут требовать использования многократного решения задач автоматической группировки).

Часть алгоритмов автоматической группировки, например, метод минимальной энтропии [182] используют в своем описании определения из информационной теории. Был предложен метод информационного узкого места или бутылочного горлышка (IBC–Information Bottleneck Clustering) [195, 199, 193], представляющий собой обобщение теории зависимости искажения информации от

скорости передачи и оперирующий представлениями, близкими для алгоритмов сжатия данных с потерями. Данный метод применяется, например, для группировки документов по ключевым словам [193]. Работа IBC-методов начинается с группировки, при которой каждый объект представляется отдельной группой (кластером) с дальнейшим последовательным исключением одной группы так, чтобы для текущего числа групп максимизировать меру сходства, измеряемую в понятиях информационной теории. Реализация данных методов требует значительных вычислительных ресурсов.

Несмотря на то, что существуют методы решения задачи автоматической группировки на основе модели k-средних [163], претендующие на нахождение глобально оптимального решения (практически неприменимые к очень большим задачам и все же не гарантирующие точное решение), все-таки основное направление современных исследований состоит в развитии эвристических алгоритмов, находящих субоптимальные решения, но при этом близкие к истинному оптимуму задачи.

В общем случае задача четкой кластеризации имеет нелинейную целевую функцию[11] (исключение составляют частные случаи, не представляющие практического интереса), со множеством локальных экстремумов. Для получения точных решений таких задач требуется решение сложной комбинаторной задачи для поиска оптимального варианта разбиения. Получение решения любой такой задачи возможно полным перебором. При этом число различных вариантов разделения N объектов на k групп оценивается как [164]:

$$M(N,k)=\frac{\sum_{i=1}^{k}(-1)^{i}\binom{k}{i}(k-i)^{N}}{k!}.$$

В алгоритмах полного перебора вычислительная сложность экспоненциально зависит от объема данных (числа группируемых объектов N) в задаче.

В алгоритмах кластерного анализа, считающихся классическими (например, в стандартной процедуре k-средних) выполняется направленный поиск в относительно небольшом подмножестве пространства возможных решений, не

гарантирующий нахождение строго оптимального решения с использованием различных ограничений (например, на число и форму получаемых групп). Эти алгоритмы реализуют локальный поиск, последовательно улучшая заранее известный промежуточный результат (но в строгом смысле их нельзя отнести к оптимизационным методам локального поиска ввиду того, что поиск следующего решения осуществляется не обязательно в окрестности предыдущего), и, таким образом, наблюдается зависимость результата этих алгоритмов от выбранного начального решения. Для поиска оптимального решения возможна организация работы в виде мультистарта данных методов с рандомизированными процедурами для выбора начальных решений в сочетании с рандомизированной процедурой локального поиска [202, 19] или более сложные подходы [11], в том числе основанные на идеях, позаимствованных из живой природы. К таким подходам относятся, в том числе, генетические алгоритмы [129], а также более широкий класс эволюционных алгоритмов, методы имитации отжига [155], нейронные сети [156] и т.д. Экспериментально исследовано преимущество эволюционных алгоритмов перед классическими алгоритмами [157]. В эволюционных алгоритмах применяются принципы и терминология, связанная с природной эволюцией. К примеру, понятие популяции – набора промежуточных решений (для нашего случая различных вариантов группировки) или «хромосом» («особей») и специальных операторов – процедуры селекции – отбора «особей» (например, отбора лучших или случайного отбора), процедуры рекомбинации (скрещивания, кроссинговера) и процедуры мутации – случайной замены некоторых элементов решения на другие, позволяющих из одной или нескольких (обычно – двух) родительских «особей» получить одну или несколько «особей» - потомков. Предложенный в [140] генетический алгоритм выполняет имитацию адаптации популяции «особей» к некоторому аналогу природной среды. При этом в алгоритме вводится специальная мера приспособленности (fitness function, т.е. функция полезности), которая, как правило, определяется на основе целевой функции задачи. В данном исследовании понятия «целевая функция» и «функция приспособленности» приняты эквивалентными. В процессе работы алгоритма происходит последовательная

смена популяций, состоящих из N_{POP} «особей» [22]. Следует отметить, что изменение состава популяции возможно полностью или частично, размер популяции может оставаться неизменным (в общем случае будем считать его неизменным) либо меняться динамически по ходу работы алгоритма [87, 62].

Чаще всего генетический алгоритм применяют в качестве стратегии глобального поиска. В работе [41] отмечено, что многие версии генетического алгоритма характеризуются тем, что: классические методы локального поиска легко находят собственно локальные оптимумы задачи, генетический алгоритм часто получает решение в виде глобального оптимума. При этом, если глобальный оптимум найден, задача проверки найденного оптимума на глобальность также является трудной [41].

Общая последовательность выполнения типового генетического алгоритма решения задачи четкой кластеризации описывается в обзоре [11] следующим образом. Задача автоматической группировки, общая схема генетического алгоритма:

1. Случайным образом формируется популяция «особей» - группировочных решений. При этом каждый вариант группировки представляется последовательностью из N целых чисел, соответствующих номерам групп. Оценивается значение целевой функции для каждой «особи».

2. Выполняется генерация следующей популяции, с использованием так называемых эволюционных операторов. Для случайного отбора «родительских» особей используется оператор селекции. Выбираются особи, наилучшие по критерию качества. С помощью рекомбинации (скрещивания, кроссинговера, кроссовера) из отобранных «особей», представленных последовательностями номеров групп, выполняется образование новых последовательностей (для простейшего случая – перестановкой одного или нескольких сегментов в этих последовательностях). Оператор мутации – как правило, рандомизированная процедура для изменения случайным образом последовательности. Например, возможна замена в случайно выбранном элементе одного номера группы другим. Для новой популяции вычисляются значения целевой функции. Этот шаг

повторяется, пока не будет выполняться заданное условие остановки.

Используются и иные способы кодирования «особей». Например, в [189] используется кодирование в виде множества индексов группируемых объектов, выбранных в качестве центров групп в k-медоидной задаче, или p-медианной задаче на сети. Так же можно закодировать множество группируемых объектов, выбранных в качестве начальных центров кластеров, которые затем уточняются процедурой k-средних или иными процедурами локального поиска [174]. Кодирование с использованием вещественных чисел –множеством точек пространства характеристик, выбранных в качестве центров групп [92, 166] используется реже.

Чиоу и Лян [110] использовали эволюционные операторы генетического алгоритма, позволяющие находить узлы сети, наиболее подходящие в качестве элементов решения (центров), при этом остальные узлы могут становиться центрами группы, если это улучшает целевую функцию.

В [102] Бозкая и др. предложен генетический алгоритм, использующий кодирование "особей" множеством индексов узлов сети, выбираемых в качестве центра группы с использованием сразу трех операторов кроссинговера. Альп, Эркут и Дрезнер [88] предложили описание простого и достаточно быстрого генетического алгоритма. В этом алгоритме используется особая процедура рекомбинации - жадная (агломеративная) эвристика. Под «эвристиками» здесь и деле будем понимать эвристические алгоритмы (процедуры). Данная эвристика вместо перестановки последовательностей, которыми представлены родительские "особи", производит объединение родительских множеств индексов узлов, выбранных в качестве центров групп. Результатом является в общем случае недопустимое решение (дочернее решение содержит больше центров, чем требуется по условиям задачи). Следующим шагом происходит последовательное удаление лишних центров до достижения допустимого решения. Каждый раз происходит удаление того элемента решения, удаление которого дает наименьший прирост целевой функции.

Возможно использование генетических алгоритмов, в том числе

применяемых к задачам автоматической группировки и размещения, без использования методов локального поиска, но эффективность таких алгоритмов, как правило, сравнительно невысока. Для решения, построенного оператором кроссинговера, необходима частичная перестройка методами локального спуска. Это обусловлено тем, что [4] эволюционный алгоритм в этом случае сосредотачивается только на поиске на множестве локальных оптимумов, несравнимо меньшем по мощности в сравнении с множеством всех допустимых решений задачи, хотя число локальных оптимумов также может расти экспоненциально с ростом объема исходных данных. Экспериментально показано, что сосредоточения множества локальных оптимумов вблизи глобального оптимума имеют высокую вероятность [101], при этом очередное сгенерированное генетическим алгоритмом решение на основе двух локальных оптимумов имеет более высокие шансы найти глобальный оптимум [4].

Генетические, а также более широкий класс – эволюционные алгоритмы – лишь определяют общую схему организации поиска. Реализация этой схемы зависит от выбора процедур селекции, мутации и особенно – рекомбинации (кроссинговера, скрещивания), которая изначально задумывалась как простая рандомизированная процедура [140], а затем была реализована для конкретных задач оптимизации в виде иногда очень сложных по своей структуре алгоритмов [181]. Эффективность отхода от рандомизированных процедур рекомбинации в сторону поиска наилучшего способа рекомбинации доказана практикой решения многих NP-трудных задач [85, 122]. Задача построения наилучшей «особи» как потомка от двух известных «особей» - родителей рассматривается в [122]. Кодирование «особей» предусмотрено в виде последовательностей нулей и единиц. Совпадение значения цифры в родительских последовательностях приводит к тому, что она наследуется «особи» - потомку. Несовпадение цифр влечет за собой необходимость решения задачи выбора значения этой цифры, обеспечивающее потомку наилучшее значение целевой функции. Следует отметить, что потомок в первом поколении с наилучшим значением целевой функции не гарантирует таковых значений в последующих поколениях. Таким образом проявляется

«жадность» алгоритма при выборе наилучшего потомка в первом поколении. Но и такая ограниченная задача для построения наилучшего потомка от двух известных родительских «особей» оказалась NP-трудной, например, применительно к p-медианной задаче (при этом для других NP-трудных задач оптимальный потомок может быть найден за полиномиальное время). На основании этого можно сделать предположение о том, что правильный выбор процедуры кроссинговера может быть задачей не менее трудной, чем собственно нахождение оптимального решения задачи оптимизации. Под «правильным» выбором понимается выбор такой процедуры, которая позволяет получить лучший результат за ограниченное время.

В диссертации Ю.А. Кочетова [41] исследованы различные методы локального поиска, в том числе некоторые жадные алгоритмы, для дискретных задач размещения (в том числе для задачи о p-медиане). Исследуются, в частности, два вероятностных жадных рандомизированных алгоритма поиска в достаточно простых окрестностях для многостадийной задачи размещения. Отмечено, что рандомизированный характер поиска позволяет сократить погрешность результата. Один из данных алгоритмов, аналогично рассматриваемой в настоящей работе жадной агломеративной эвристике, построен на последовательном сокращении мощности множества, которым представлено промежуточное решение, при этом удаляемый из множества элемент выбирается случайно из некоторого подмножества потенциально удаляемых объектов.

Число особей, т.е. размер популяции является важным параметром генетического или любого эволюционного алгоритма. Использование чрезмерно ограниченного размера популяция алгоритма при отсутствии процедуры мутации (либо при незначительном влиянии мутации) приводит генерации одних и тех же решений (т.н. «вырождение» популяции) [173]. В [63] авторами показано, что для достижения наилучшего результата, как правило, возможно при соблюдении приблизительного равенства сменяющихся поколений «особей» и числе самих «особей». Использование вычислительно сложной процедуры кроссинговера не позволяет сделать точное предсказание времени ее выполнения, и, следовательно, числа сменяющихся поколений за установленное время. Исследование

зависимости времени достижения локального оптимума от размера популяции и размера турнира (размер турнира s – число «особей», среди которых выбираются лучшие в качестве родительских «особей», порождающих «особи» следующего поколения) выполнено в [22]. Пусть h – число неоптимальных решений в окрестности, L – минимальная вероятность достижения локального оптимума в очередном поколении, E – вероятность того, что, хотя бы одна «особь»-потомок будет по значению целевой функции не хуже родительской, r – доля «особей» популяции, участвующих в турнирной селекции. Тогда размер популяции и размер турнира оцениваются как

$$N_{POP} = 2\left\lceil \frac{1+\ln h}{L \cdot E\left(1/e^{2r}\right)} \right\rceil, s = \lceil rN_{POP} \rceil,$$

где $\lceil \cdot \rceil$ - округление вверх. .

Получение такой оценки размера популяции возможно только после решения довольно сложных задач по определению значений *h, L, E*. Оценки этих значений даются в [22] для случаев с простыми процедурами кроссинговера. Включение локального поискав процедуру кроссинговера приводит к выбору любых двух родительских «особей», дающих в результате локальный оптимум в качестве потомка. Таким образом, для достижения локального оптимума достаточно N_{POP}=2. Это никак не отражает размер популяции, необходимой для получения близких к глобальному оптимуму решений.

1.2 Постановка задачи разделения смеси распределений

Для задач автоматической группировки данных возможно применение подходов из смежной области - вероятностной диагностики (разделе математической статистики, предметом рассмотрения в которой являются проблемы обнаружения изменений вероятностных свойств данных), т.е. задач обнаружения изменения вероятностных свойств данных [14]. К таким задачам относятся: задачи о разладке, т.е. оценки момента изменения закона распределения

данных, а также задачи разделения (расщепления) смеси распределений.

Согласно классификации задач вероятностной диагностики, рассматриваемые в настоящем исследовании задачи разделения смеси вероятностных распределений могут быть классифицированы следующим образом [14]:

- по характеру объекта диагностики: как задачи вероятностной диагностики для случайных полей;

- по механизму смены состояний объекта диагностики: образуют обособленный класс задач о разделении смеси вероятностных распределений;

- по способу получения информации об объекте: апостериорный либо последовательный анализ, т.е. все данные (измерения) получены до начала работы алгоритма. В настоящем исследовании изучаются исключительно методы и алгоритмы апостериорного анализа как позволяющие находить более точные решения;

- по характеру методов вероятностной диагностики: параметрические, т.е. вероятностная модель данных, определяемая параметрами распределений.

Задачи разделения смеси распределений относятся к классу задач автоматической группировки на основе плотности распределения с неизвестными параметрами. Целью решения такой рода задачи является нахождение параметров этих распределений. Критерием нахождения искомых параметров является логарифмическая функция правдоподобия.

Решение задач автоматической группировки подразумевает использование данных различного происхождения. Данные могут быть порождены различными видами распределений вероятностей случайной величины. В нашей работе мы исходили из того, что весь объем данных порожден смесью нескольких многомерных распределений. Эти распределения подчиняются нормальному или близкому к нормальному распределению (распределение Лапласа, экспоненциальное распределение). Для наших практических данных хорошая апроксимируемость нормальным распределением показана в Главе 3.

Общая постановка задачи разделения смеси распределений состоит в

следующем:

Пусть плотность распределения на множестве X имеет вид смеси k распределений (предполагаем, что распределения гауссовы) [39]:

$$\rho(x)=\sum_{j=1}^{k}\alpha_j\rho_j(x),\ \sum_{j=1}^{k}\alpha_j=1,\ \alpha_j\geq 0,$$

где $\rho_j(x)$ - функция правдоподобия j-ой компоненты смеси, α_j- ее априорная вероятность («вес» в составе смеси).

Приведем краткое описание некоторых видов распределений [2]:

Дано: N векторов d-мерных данных $X_i=\left(x_{i,1},\dots,x_{i,d}\right)^T\ ,i=\overline{1,N}$.

Многомерное гауссово распределение

Существует вектор $\mu\in R^d$ и неотрицательно определённая симметричная ковариационная матрица

$$\Sigma=\begin{pmatrix}\sigma_1^2=\sigma_{1,1} & \sigma_{1,2} & \dots & \sigma_{1,d}\\ \sigma_{2,1} & \sigma_2^2=\sigma_{2,2} & \dots & \sigma_{2,d}\\ \vdots & \vdots & \ddots & 0\\ \sigma_{d,1} & \sigma_{d,2} & \dots & \sigma_d^2=\sigma_{d,d}\end{pmatrix}p$$

размерности $d\times d$, такие что плотность вероятности вектора X имеет вид:

$$f(X)=\frac{\alpha}{\sqrt{(2\pi)^d|\Sigma|}}exp(-\frac{1}{2}(X-\mu)^T\Sigma^{-1}(X-\mu)),$$

где $|\ \Sigma\ |$— определитель матрицы Σ, а Σ^{-1}— матрица обратная к Σ.Компоненты вектора μ вычисляются по отдельности для каждого измерения: $\mu_j=\frac{1}{N}\sum_{i=1}^{N}x_{i,j}$

Здесь $x_{i,j}$–й компонент (j-е измерение) i-го вектора данных.

Компоненты ковариационной матрицы вычисляются так:

$$\sigma_{j,k}=\sigma_{k,j}=\frac{1}{N}\sum_{i=1}^{N}\left(x_{i,j}-\mu_j\right)\left(x_{i,k}-\mu_k\right).$$

Многомерное гауссово распределение с независимыми (некоррелированными) признаками (измерениями)

То же, что в предыдущем случае, но не требует работы с матрицами.

Компоненты вектора μ как в предыдущем случае.

Ковариационная матрица диагональная (поэтому может быть заменена вектором), состоит из дисперсий по каждому измерению:

$$\Sigma = \begin{pmatrix} \sigma_1^2 & 0 & \dots & 0 \\ 0 & \sigma_2^2 & \dots & 0 \\ \vdots & \vdots & \ddots & 0 \\ 0 & 0 & \dots & \sigma_d^2 \end{pmatrix}.$$

Дисперсии рассчитываются независимо: $\sigma_j^2 = \frac{1}{N}\sum_{i=1}^{N} \left(x_{i,j} - \mu_j\right)^2$.

Плотность распределения вычисляется как произведение плотностей по каждому измерению:

$f(X) = \prod_{j=1}^{d} f_j(x_j) = \prod_{j=1}^{d} \frac{1}{\sigma_j\sqrt{(2\pi)}} exp(-\frac{1}{2}\left(x_j - \mu_j\right)^2/\sigma_j^2)$, где – j-й компонент вектора X.

Сферическое гауссово распределение

То же, что в предыдущем случае, но среднеквадратичные отклонения по всем измерениям равны между собой: $\sigma_j^2 = \sigma^2 = \frac{1}{Nd}\sum_{j=1}^{d}\sum_{i=1}^{N} \left(x_{i,j} - \mu_j\right)^2$. Вместо ковариационной матрицы имеем скаляр.

Любое многомерное распределение с независимыми признаками (измерениями)

Плотность распределения вычисляется как произведение плотностей по каждому измерению:

$f(X) = \prod_{j=1}^{d} f_j(x_j)$,

где – j-й компонент вектора X.

В этом случае $f_j(x_j)$– плотность одномерного распределения (например, нормального, Лапласа, экспоненциального, Релея и т.д.). Параметры каждого из одномерных распределений определяются независимо.

Экспоненциальное распределение

Плотность в точке x оценивается как $f(x) = \lambda\, exp(-\lambda x)$.

Это распределение определено только для неотрицательных значений *x*.

Единственный параметр λ связан с математическим ожиданием μ следующим образом:$\mu = 1/\lambda$.

Он же связан со среднеквадратичным отклонением так:$\sigma^2 = \frac{1}{\lambda^2}$.

Распределение Лапласа (оно же – двойное экспоненциальное)

Плотность в точке *x* оценивается как$f(x) = \frac{\lambda}{2} exp(-\lambda|x - \mu|)$,

коэффициент λ связан со среднеквадратичным отклонением следующим образом: $\sigma^2 = \frac{2}{\lambda^2}$.

Пусть функции правдоподобия принадлежат параметрическому семейству распределений $\varphi(x;\theta)$ и отличаются только значениями параметра $\rho_j(x) = \varphi(x;\theta_j)$.

Задача разделения смеси (нечеткой кластеризации) заключается в том, чтобы, имея выборку $\boldsymbol{X}^m$ случайных и независимых наблюдений из смеси $\rho(x)$, зная число *k* и функцию φ, оценить вектор параметров $\Theta = (\alpha_1,...,\alpha_k,\theta_1,...,\theta_k)$.

Здесь $\theta_1,...,\theta_k$ – собственно параметры распределений, а $\alpha_1,...,\alpha_k$ – их априорные вероятности, или, иными словами, весовые коэффициенты в смеси распределений, сумма которых равна 1.

Эти параметры отыскиваются путем решения задачи максимизации логарифма полного правдоподобия:

$$Q(\Theta) = ln\prod_{i=1}^{m}\rho(x_i) = \sum_{i=1}^{m} ln\sum_{j=1}^{k}\alpha_j\rho_j(x_i) \to \max_{\Theta}.$$

Характер прикладных задач, решаемых в настоящем исследовании, обусловил то, что данная работа в основном связана с задачами разделения многомерных гауссовых(нормальных) распределений и их частными случаями –

некоррелированным и сферическим, а также распределением Лапласа.

1.3 Основные методы решения задачи разделения смеси распределений

Для разделения смесей вероятностных распределений, начиная с конца 60-х годов прошлого столетия, предложено достаточное количество подходов [26]. В процессе такого разделения выделяют две основные задачи: оценивание параметров смесей (1) и построение решающего правила разделения объектов между смесями (2). Порядок решения этих задач может быть различным. Для решения задач в порядке (1) – (2) необходимо обладать информацией о типах распределений. Для этого можно использовать известные классические методы, такие как метод моментов, метод максимального правдоподобия, метод наименьших квадратов и др. В случаях, когда указанная информация отсутствует или ее использование затруднено (например, из-за вычислительной трудности) то используют последовательность решения (2) – (1), т.е. решают задачу кластеризации – «обучение без учителя».

Обобщенный метод моментов, являющийся обобщением классического метода моментов бы предложен в Хансеном 1982 г. [132]. Метод применяется для оценки параметров в статистических полупараметрических моделях, когда искомые параметры конечномерны, тогда как форма функции распределения данных может быть неизвестна. Таким образом, оценка функции правдоподобия неприменима. Суть данного метода заключается в том, что к выборочным или заданным моментам смеси приравниваются теоретические моменты смеси, выраженные через теоретические моменты составляющих смесь распределений.

Разделение любого конечного числа экспоненциальных распределений показано в [158]. Исследованы некоторые свойства выборочных оценок параметров: для наилучших асимптотических нормальных оценок [99] показана состоятельность и неэффективность оценок параметров по м.м. [45, 115]. Асимптотическая нормальность оценок доказана в [148, 147, 192]. Сделано общее заключение [115] о применимости моментных оценок в одномерном случае и при

больших расстояниях между популяциями моментные оценки, хотя и не являются эффективными, могут использоваться для практических расчетов. Решение системы уравнений по м.м. для нахождения оценок параметров неоднозначно [139, 150]. Для многомерных задач со смесями нормальных распределений рассмотрено в [16, 148].

Применение метода максимума правдоподобия для отыскания параметров смесей широко описано в литературе. Алгоритмы данного итерационного метода, приводящего функцию правдоподобия к максимуму, вычисляют по некоторым значениям параметров новые (более правдоподобные). Доказаны теоремы о сходимости алгоритмов к оценкам максимального правдоподобия независимо то начальных значений параметров. Уравнения правдоподобия преобразуются к виду

$$\theta_j = \Phi(\theta_{j-1}),$$

где θ – искомый набор параметров, а *j*–номер итерации.

Метод для конечной смеси нормальных распределений дан в [84, 94, 115, 138, 148]. Возможно применение различных итерационных схем, например, метод скорейшего спуска и метод Ньютона [138]. Показано, что для объема выборки больше 200 метод скорейшего спуска всегда увеличивает значение функции правдоподобия и хорошо работает при малом числе данных; метод Ньютона хорош для больших объемов данных и имеет результатом более близкие к истинным оценки параметров распределений. Также автор указывает на очень возрастание асимптотических оценок дисперсий для весов с уменьшением расстояния между средними.

Применение метода максимального правдоподобия показано для нахождения параметров смесей экспоненциальных распределений [137], смеси двух гамма-распределений [147].

Исследовано применение м.м.п. для случая с малым объемом выборки [136]. Подсчитаны асимптотические дисперсии для оценок параметров [115, 138, 137, 191].

В сравнительном анализе [126, 127] м.м. и м.м.п. предпочтение отдается м.м.п. как наиболее эффективному. Для многомерных задач и при нормальном

характере распределений смеси это преимущество возрастает.

В работах [115, 206, 148, 187] предложены разные процедуры применения м.м.п.

В [148] функция правдоподобия описывается в другом виде:

$$\lambda = (2\pi)^{-N\cdot p/2}\left|\Sigma\right|^{-N/2} exp\left[-\sum_{i=1}^{2}\sum_{r=1}^{N}\theta_{ir}(x_r-\mu_i)^T\Sigma^{-1}(x_r-\mu_i)/2\right]$$

и максимизируется надлежащим выбором θ_{ir}. Параметры μ_1, μ_2, Σ выражаются через $\hat{\theta}_{ir}$ (значение θ_{ir}, при котором функция правдоподобия максимальна) равенствами:

$$\mu_i = \frac{\sum_r \hat{\theta}_{ir}x_r}{\sum\hat{\theta}_{ir}},\ i=1,2,\ \sum = \sum_i\sum_{r=1}^{N}\hat{\theta}_{ir}(x_r-\mu_i)(x_r-\mu_i)^T/N.$$

Для нахождения $\hat{\theta}_{ir}$ предложена итерационная процедура, сходимость которой не доказана.

В работе [187] рассмотрена следующая модель. Кроме $\mu_1, \ldots, \mu_M, \Sigma_1, \ldots, \Sigma_M$, ищется параметр группирования $\gamma=(\gamma_1, \ldots, \gamma_N)$, где γ_g равно g, если элемент выборки взят из g-ой совокупности. Если $\theta=(\gamma, \mu_1, \ldots, \mu_M, \Sigma_1, \ldots, \Sigma_M)$, то логарифмическая функция правдоподобия определяется как

$$l(\theta) = -\frac{1}{2}\cdot\sum_{g=1}^{M}\left(\sum_{G_g}(y_i-\mu_g)^T\Sigma^{-1}(y_i-\mu_g)+n_g\cdot log\left|\Sigma_g\right|\right),$$

где G_g – множество элементов выборки, отнесенных в g-ую группу, n_g–число элементов в G_g, $|\Sigma_g|=\det\Sigma_g$. Проблема состоит в оценивании N-мерного вектора γ и, как следствие, группы $G_1, G_2, \ldots, G_M$. Оценка $\hat{\gamma}$ по методу максимального правдоподобия γ находится при минимизации

$$\sum_{g=1}^{M}\left\{trW_{g_y}\Sigma^{-1}+n_g\cdot log\left|\Sigma_g\right|\right\},\quad \text{где}\quad W_{gy}=\sum_{G_g}(y_i-\bar{y}_g)(y_i-\bar{y}_g)^T.$$ Далее, если $\Sigma_g=\Sigma(g=1, \ldots, M)$ и, считая Σ известной (или имеем достаточно большую выборку для получения «хорошей» оценки для Σ), $\hat{\gamma}$ находится при минимизации

$$tr(W_y\cdot\Sigma^{-1}) = tr\left(\sum_g W_{gy}\cdot\Sigma^{-1}\right) = \sum_{g=1}^{M}\sum_{G_g}(y_i-\bar{y}_g)^T\Sigma^{-1}(y_i-\bar{y}_g)^T,$$

т.е. минимизируется общая внутригрупповая сумма квадратов. Это один из известных экстремальных подходов к разделению объектов на группы. Если Σ не является известной, то $\hat{\gamma}$ есть тот параметр группирования, который минимизирует W_y; этот подход также широко применяется в задачах кластеризации. Далее снимается ограничение на равенство ковариационных матриц составляющих группы, и тогда $\hat{\gamma}$ ищется при минимизации $\prod_{g=1}^{M}\left|W_{gy}\right|^{n_g}$ (необходимо иметь по крайней мере p+1 наблюдение из каждой группы, где p–размерность пространства).

В работах приводятся числовые результаты и показаны возможности применения указанных подходов. Указывается на необходимость дальнейшего исследования свойств полученных оценок и расширения методики для случаев, когда не задано количество искомых распределений в смеси.

В работах [47, 126] сравниваются м.м., м.м.п. и метод минимума χ^2 для разделения смеси распределений. Система уравнений для метода максимального правдоподобия труднореализуема ввиду того, что имеются решения, соответствующие седловым точкам поверхности L (функция правдоподобия) в пространстве параметров. Также с ростом Nсреднеквадратичная погрешность оценок уменьшается со скоростью $\sqrt{N}$, а объём вычислений растет пропорционально N–в то время как метод минимума χ^2близкий по универсальности и эффективности методу максимального правдоподобия не требует увеличения объема вычислений.

Использование различных типов выборок смеси нормальных распределений, моделируемых по методу Монте-Карло и использование дополнительной информации об одном из распределений показано в [141, 118].

Метод наименьших квадратов можно использовать для определения весов (т.е. априорных вероятностей) распределений смеси для случаев, когда параметры этих распределений известны [93, 112, 111, 178]. При этом веса искомых распределений находятся из условия минимума интеграла квадратичного отклонения заданного распределения смеси от распределения, отвечающего модели смеси с варьируемыми значениями весов. В работе [178] предлагается

способ построения функций для оценки смешивающего распределения. Оценка весов составляющих распределений, применимая к непрерывным и дискретным смесям приводится в [112]; показано, что описанный метод хорошо работает на малых объемах данных. Дальнейшее развитие этого метода, позволяющего находить оценки не только для весов, но и для параметров распределений со скоростью сходимости $O(N^{-1/2})$ приводится в [111].

Для задач, в которых неизвестны только веса составляющих распределений могут применяться *методы с использованием элементы выборки для оценивания параметров смеси*. Минимаксная оценка [100] минимизирует максимум функции риска, выраженной через взвешенную сумму диагональных элементов обратной информационной матрицы. Состоятельная байесовская оценка приводится для дискретных [170] и непрерывных [146] смесей. Предложены оценки в виде различных типов функций: ступенчатой с конечным числом ступенек [116] и кусочно-полиномиальными дугами [179]. Общий алгоритм конструирования состоятельной оценки всех идентифицируемых семейств смесей распределений приведен в [207].

Предложен способ построения критерия относительно веса распределения для 2-х распределений. Например [95], при гипотезе $p=0$ выборочное среднее распределено нормально; при гипотезе $0 < p < 1$ $\bar{X}$ является смесью $(n+1)$ -го нормального распределения с общей дисперсией $\frac{\sigma^2}{n}$ и со средним $\mu_k = \frac{k}{n}\mu_1 + \left(1 - \frac{k}{n}\right)\mu_2$. В [109] оценка $\hat{p}$ весового коэффициента смеси многомерных распределений для p предлагается в виде отношений линейных комбинаций меток рангов наблюдений. Применение графических методов разделения смеси на составляющие показано в [130]. Предложен [208] алгоритм стохастической аппроксимации для смеси одномерных нормальных функций распределения. В [149] предлагается метод оценки параметров p_i и μ_i, $i=1, \ldots, r$ для плотности вероятности вида $f(x) = X_{\chi}(x)\sum_{l} p_i \, exp\{ B(\mu_i)x + C(x) + D(\mu_i)\}$,

где $X_{\chi}(x)$ - характеристическая функция заданного интервала, *B, C, D* – известные функции. Показана состоятельность и асимптотическая нормальность полученных оценок. Применена общая теория для оценки смесей двух биномиальных и смеси двух экспоненциальных распределений. Для однопараметрических смесей задача разделения смесей предложена [197] с помощью решения дифференциального уравнения. Использование преобразования Фурье для разделения двух нормальных совокупностей предлагается в [169, 194].

Упомянутые работы основываются на построения некоторых алгоритмов для нахождения оценок параметров смеси. Данные работы не позволяют получить аналитического выражения оценок для случаев, когда нельзя ввести упрощающие задачу некоторые предположения о свойствах распределений смеси. В приведенных задачах число распределений смеси фиксировано. Не дается рекомендаций для решения задачи для случаев, когда количество компонент смеси неизвестно.

Описанные методы используют данные выборки только на первом этапе расчетов для оценивания параметров смеси, и все дальнейшие вычисления производятся только с полученными статистиками.

В задачах кластеризации (т.е. объединении объектов в однородные группы) элементы выборки используются на протяжении всей вычислительной процедуры. Описание постановок задач кластеризации дано в [2]. Сравнение некоторых методов дается в [180].

Для определения оптимальности разделения объектов на группы любым методом группировки объектов необходимо введение критерия. Часто в качестве такого критерия используют условие минимума или максимума какой-либо функции взаимной близости между элементами смеси. Выделяют три основных способа количественного выражения сходства элементов – коэффициентами правдоподобия для иерархической классификации, коэффициентами корреляции для задач факторного анализа и различными метрическими показателями расстояния. Обсуждение мер близости и вопросы выбора критериев качества приведены в [2].

Для нахождения оптимального разбиения заданной совокупности элементов на однородные группы возможно применение идей дискриминантного анализа [167, 201]. В этом случае производят последовательное выделение одного или группы признаков, по которому происходит разбиение на две группы с дальнейшим ранжированием по этим признакам и построения дискриминантной функции разделения по всем признакам. Далее происходит новое разбиение на группы с учетом результатов ранжирования и всю процедуру выполняют заново. Лучший результат фиксируется. Таким образом производится би-кластеризация. Разделение на несколько составляющих совокупностей производится методом дихотомий. Многократное применение процесса дихотомии для разделения на любое число групп не обеспечивает оптимума разбиения. Для оптимизации процесса необходимо ввести критерий объединения [20, 165]. Таким образом, введением такого критерия объединения задается правило остановки иерархической группировки. В работах [125, 187] для определенного числа классов предлагается использовать критерий минимизации локальной энтропии. Но для общего случая вопрос определения числа групп не решен.

Для подходов, использующих выборки, актуален вопрос подтверждения статистической значимости найденного оптимального разделения совокупностей. Этот вопрос был рассмотрен в [121, 185, 201].

Наиболее популярным методом численного решения задачи разделения смеси распределений является EM-алгоритм [117] (Expectation Maximization – максимизация математического ожидания) и его модификации. Результатом работы EM-алгоритма является нахождение параметров каждого из распределений и их априорных вероятностей.

EM-алгоритм для разделения смеси, например, сферических гауссовых распределений может быть описан следующим образом. Имеется набор данных $S \subset R^n$, EM-алгоритм для смеси k нормальных распределений с общей сферической матрицей ковариации стартует с начальными стартовыми значениями параметров $\mu_i\langle 0\rangle$, $\alpha_i\langle 0\rangle$, $\sigma_i\langle 0\rangle$ (вектор математического ожидания i-го распределения, весовой коэффициент i-го распределения в смеси и его среднеквадратическое отклонение

соответственно), которые в дальнейшем обновляются в соответствии со следующей двухшаговой процедурой (здесь t – номер итерации).

Алгоритм 1.1 EM-алгоритм.

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 $x \in S$ 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. $\alpha_i^{\langle t+1 \rangle} = \frac{1}{N} \sum_{x \in S} p_i^{\langle t+1 \rangle}(x); \mu_i^{\langle t+1 \rangle} = \frac{\sum_{x \in S} x p_i^{\langle t+1 \rangle}(x)}{N \alpha_i^{\langle t+1 \rangle}};$

$$\sigma_i^{\langle t+1 \rangle 2} = \frac{1}{d} \sum_{x \in S} \left\| x - \mu_i^{\langle t+1 \rangle} \right\|^2 p_i^{\langle t+1 \rangle}(x) \qquad (1.1)$$

Повторять с шага 1.

Останов алгоритма, если нет приращения целевой функции правдоподобия

$$L^{<t+1>} = \sum_{x \in S} \sum_{i=1}^{k} ln(\tau_i p_i^{\langle t+1 \rangle}(x)). \qquad (1.2)$$

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 (1.1) 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. Алгоритм может быть адаптирован практически для любого вида распределений, заранее известного с точностью до числовых параметров.

EM-алгоритм достаточно хорошо исследован, получены различные модификации, улучшающие его сходимость и устойчивость к возмущениям во входных данных.

Основные модификации EM-алгоритма:

- Классический алгоритм (пример приведен выше);

- Стохастический SEM-алгоритм [103, 106, 107]. Выполняется целенаправленное «встряхивание» выборок на каждой итерации для уменьшения эффекта, при котором алгоритм «скатывается» к ближайшему локальному максимуму;

- Классификационный CEM-алгоритм [105, 108]. Алгоритм выполняет детерминированное разделение на кластеры. Объект четко относится к тому кластеру (распределению), вероятность принадлежности к которому максимальна из полученных;

Стохастический алгоритм дает более стабильные результаты. Но они не являются оптимальными.

Алгоритмы популярны, являются де-факто стандартом разделения смеси распределений.

Но алгоритм имеет ряд существенных недостатков:

- требует задания предполагаемого числа распределений в смеси

- необходимо задание начального решения: алгоритм является (не в строгом смысле) алгоритмом локального поиска

К преимуществам EM относят:

- не требуется использования метрик;

- его можно комбинировать с другими алгоритмами обработки данных;

- алгоритм работает с малыми данными.

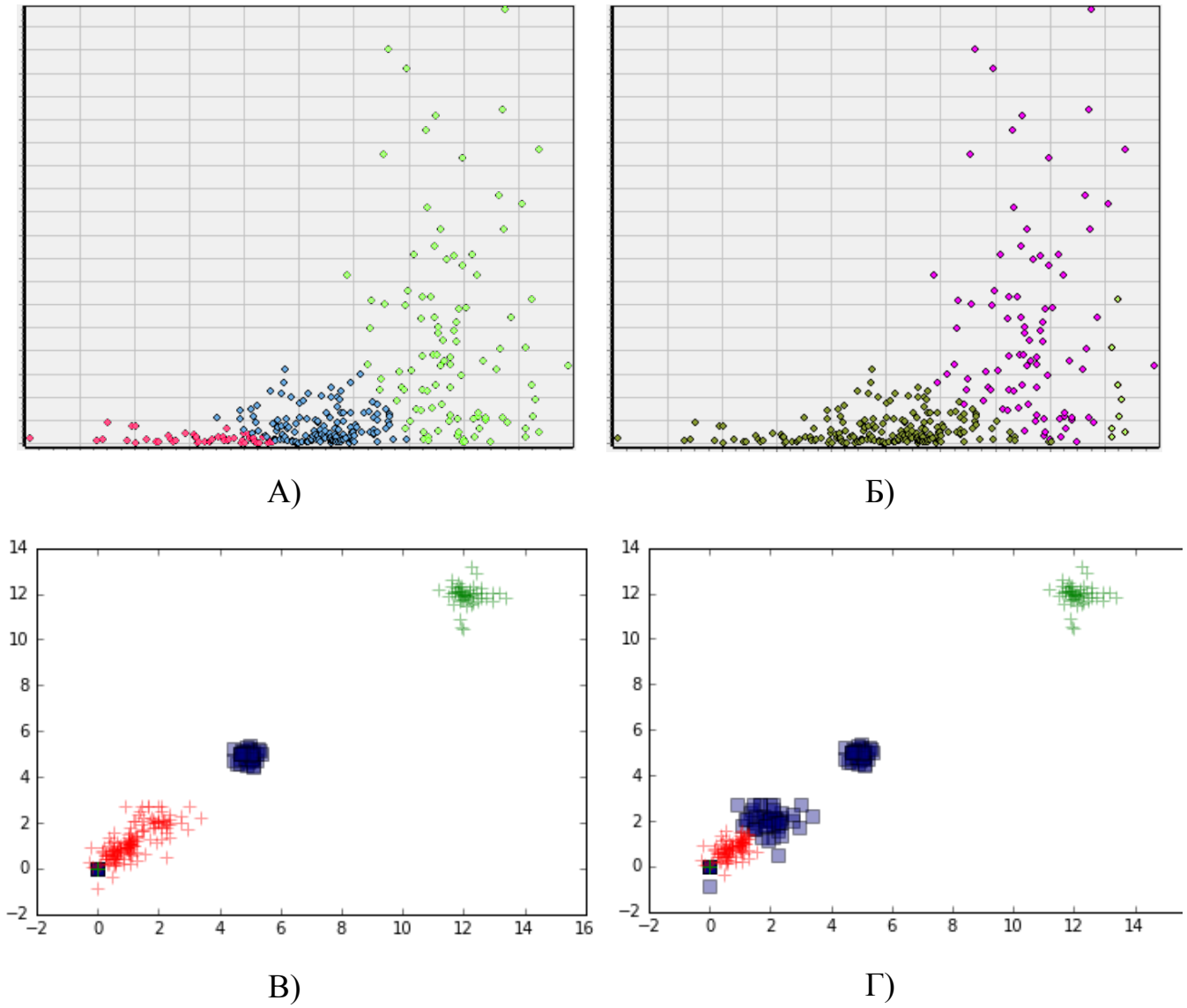


Рисунок 1.1–Пример зависимости результата от начального решения

На рис. 1.1 показан пример зависимости результата четырех независимых попыток разбиения объектов на три кластера от начального решения.

На рис. А и В показаны результаты разбиения объектов на группы, соответствующие максимальному значению функции правдоподобия. Для случаев Б и Г результаты работы оказались не столь точными (значение максимизируемой функции правдоподобия меньше, чем для результатов А и В).

В многомерных случаях и при большом числе кластеров эта зависимость усугубляется, что подтверждается результатами расчетов в Главе 2. В алгоритмах для задачи k-средних эта неустойчивость проявляется не так сильно.

1.4 Метод жадных эвристик для задач автоматической группировки объектов

Для задач автоматической группировки на моделях теории размещения Л.А. Казаковцевым и А.Н. Антамошкиным предложен метод жадных эвристик [154]. Алгоритмы этого метода, в основном, рандомизированы, но при этом они позволяют получать для практических задач результаты, которые трудно существенно улучшить другими методами за сопоставимое время.

При этом эти результаты достаточно стабильны, т.е. рандомизированные алгоритмы дают при перезапусках очень близкие результаты.

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

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

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

$$\arg\min\ F(X_1,...,X_k) = \sum_{i=1}^{N} \min_{j\in\{1,k\}} L(X_j, A_i). \qquad (1.3)$$

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

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 1.2.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 и 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 (набор параметров приведен для случая сферических гауссовых распределений). В несколько меньшей степени зависимость от начального решения прослеживается и для других модификаций EM-алгоритма (например, SEM).

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.

На рис.1.2 представлена общая схема организации трех циклов метода жадных эвристик Л.А.Казаковцева и А.Н.Антамошкина. На рис.1.3 показана схема совместимости компонентов метода жадных эвристик и варианты применения метода.

Как видно из рис. 1.2 и 1.3, метод способен решать разнообразные задачи автоматической группировки, основанные на моделях теории размещения, т.е. основанные на поиске центров или центроидов кластеров. 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 задач.

В настоящей работе в качестве основной рабочей гипотезы выдвигается следующее утверждение: подходы к повышению точности и устойчивости результата метода *k*-средних можно адаптировать для задач разделения смеси распределений.

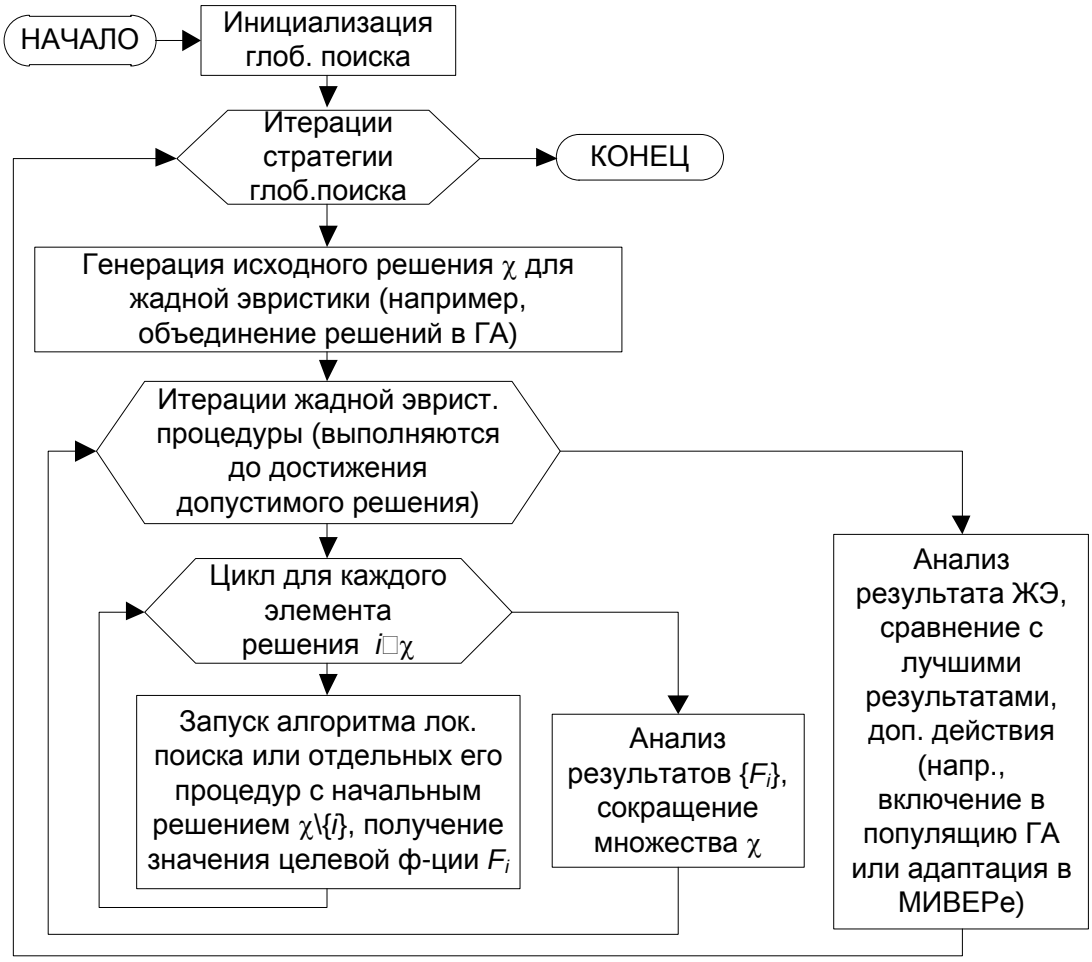


Рисунок 1.2 – Общая схема алгоритма метода жадных эвристик

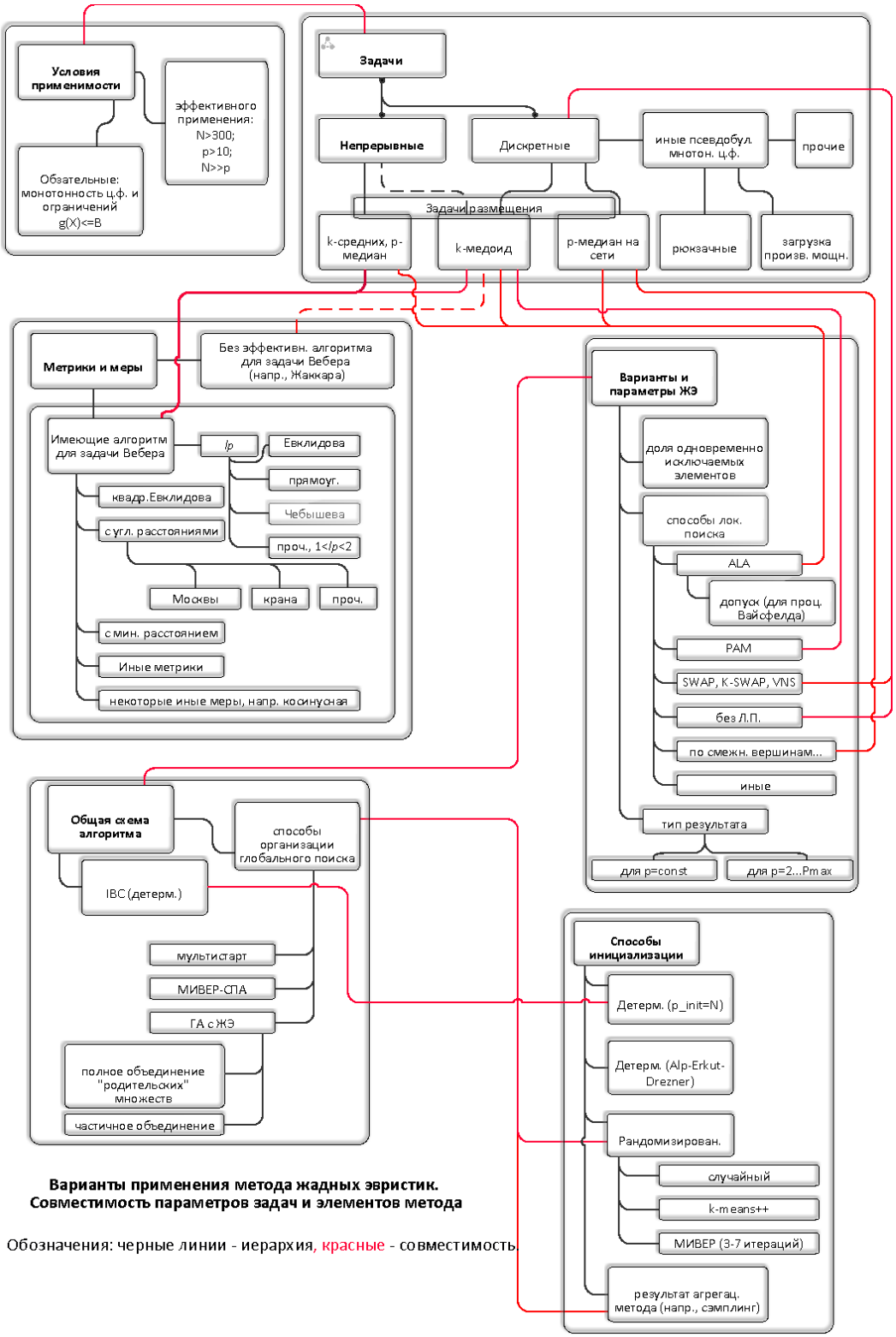


Рисунок 1.3 – Схема совместимости компонентов метода жадных эвристик

Выводы к Главе 1

Анализ известных методов автоматической группировки, основанных на модели разделения смеси распределений, выявил, что, несмотря на имеющийся широкий набор известных алгоритмов решения задачи, 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 (ALA и EM).

Таким образом, в Главе 1 показано, что решение задачи повышения точности и устойчивости методов локального поиска является актуальной задачей. Подтверждение выдвинутой в работе гипотезы о применимости подходов к повышению точности и устойчивости результата метода k-средних и возможности адаптировать эти методы для алгоритма локального поиска на основе модели разделения смеси распределений (EM-алгоритма) приведено в Главе 2.

ГЛАВА 2. АЛГОРИТМЫ МЕТОДА ЖАДНЫХ ЭВРИСТИК ДЛЯ ЗАДАЧ РАЗДЕЛЕНИЯ СМЕСИ РАСПРЕДЕЛЕНИЙ

2.1 Известные алгоритмы: EM-алгоритм и его модификации

В настоящее время существует большое количество методов кластеризации и классификации данных [37, 82, 39]. Обзор наиболее часто используемых на практике методов приведен в [37]. Как упоминалось в Главе 1, в число самых популярных методов входит и EM-алгоритм (Expectation Maximization – максимизация математического ожидания), который успешно применяется для статистических задач, связанных с анализом неполных данных, когда некоторые статистические данные отсутствуют либо для случаев, когда функция правдоподобия имеет вид, не допускающий удобных методов исследования, но допускающий серьезные упрощения при введении дополнительных «ненаблюдаемых» («скрытых») величин [28]. Именно такая постановка задачи (кластеризация многомерных данных нормального распределения со скрытыми данными) используется нами для решения задач разделения ЭРИ по производственным партиям исходного сырья в Главе 3.

Пусть плотность распределения на множестве X имеет вид смеси k распределений (предполагаем, что распределения гауссовы) :

$$\rho(x)=\sum_{j=1}^{k}\alpha_j\rho_j(x),\ \sum_{j=1}^{k}\alpha_j=1,\ \alpha_j\geq 0,$$

где $\rho_j(x)$ - функция правдоподобия j-ой компоненты смеси, α_j- ее априорная вероятность.

Пусть функции правдоподобия принадлежат параметрическому семейству распределений $\varphi(x;\theta)$ и отличаются только значениями параметра $\rho_j(x)=\varphi(x;\theta_j)$.

Задача разделения смеси (нечеткой кластеризации) заключается в том, чтобы, имея выборку X^m случайных и независимых наблюдений из смеси $\rho(x)$, зная

число k и функцию φ, оценить вектор параметров $\Theta = (\alpha_1,...,\alpha_k,\theta_1,...,\theta_k)$.

Идея алгоритма заключается в следующем. Искусственно вводится вспомогательный вектор скрытых переменных G, обладающий двумя замечательными свойствами:

- С одной стороны, он может быть вычислен, если известны значения вектора параметров Θ;

- С другой стороны, поиск максимума правдоподобия сильно упрощается, если известны значения скрытых переменных.

ЕМ-алгоритм состоит из итерационного повторения двух шагов:

На Е-шаге вычисляется ожидаемое значение (expectation) вектора скрытых переменных G по текущему приближению вектора параметров Θ. Обозначим через $\rho(x;\theta_j)$ плотность вероятности того, что объект xполучен из j-ой компоненты смеси. По формуле условной вероятности

$$\rho(x,\theta_j) = \rho(x)P(\theta_j \mid x) = \alpha_j \rho_j(x).$$

Введем обозначение $g_{ij} \equiv P(\theta_j \mid x_i)$. Это неизвестная апостериорная вероятность того, что обучающий объект x_i получен из j-ой компоненты смеси. Эти величины используются в качестве скрытых переменных.

$\sum_{j=1}^{k} g_{ij} = 1$, для любого $i = 1,..,m$. т.к. имеет смысл полной вероятности принадлежать объекту x_iодной из k компонент смеси. Из формулы Байеса

$$g_{ij} = \frac{\alpha_j \rho_j(x_i)}{\sum_{s=1}^{k} \alpha_s \rho_s(x_i)} \text{ для всех } i, j.$$

На М-шаге решается задача максимизации правдоподобия (maximization) и находится следующее приближение вектора Θ по текущим значениям векторов G и Θ.

Будем максимизировать логарифм полного правдоподобия:

$$Q(\Theta) = ln\prod_{i=1}^{m} \rho(x_i) = \sum_{i=1}^{m} ln\sum_{j=1}^{k} \alpha_j \rho_j(x_i) \to \max_{\Theta}.$$

Решая оптимизационную задачу Лагранжа с ограничением α_j, находим:

$$\alpha_j = \frac{1}{m}\sum_{i=1}^{m} g_{ij}, j = 1,..,k$$

$$\theta_j = arg\max_{\theta} \sum_{i=1}^{m} g_{ij} \ln \varphi(x_i, \theta), j = 1,..,k .$$

Таким образом, М-шаг сводится к вычислению весов компонент α_j как средних арифметических и оцениванию параметров θ_j путем решения *k* независимых оптимизационных задач. Отметим, что разделение переменных оказалось возможным благодаря удачному введению скрытых переменных.

Итерационный процесс останавливается в соответствии с заранее согласованным критерием остановки (заранее выбранная метрика $\rho(\theta_1, \theta_2)$ и число ε). Процесс останавливается на *m*-ом шаге, если $\rho(\theta^{(m)}, \theta^{(m-1)}) < \varepsilon$.

В нашей практической задаче выделения однородных партий электрорадиоизделий [76] алгоритм на основе метода жадных эвристик и модели *k*-средних [27], применяемый в настоящее время на практике для разбиения выборки электронных компонентов на однородные производственные партии, не позволяет определять количество *k* компонентов смеси (количество кластеров). Эта величина должна задаваться перед началом работы алгоритма либо должна решаться серия задач с различным предполагаемым числом кластеров.

К преимуществам EM-алгоритма можно отнести [39]:

- не требуется введение метрик;

- возможность использования совместно с другими алгоритмами обработки данных;

- работа на малых объемах данных;

- позволяет выделять шумовые выбросы и разреженный фон [82];

- разработан ряд модификаций основного алгоритма [28]: так, для частичного устранения указанных недостатков разработан ряд модификаций основного алгоритма (медианные модификации, SEM-алгоритм, CEM-алгоритм, MCEM и SAEM-алгоритмы).

Так, при применении более распространенной модели *k*-средних к задаче выделения однородных производственных партий электрорадиоизделий выбор метрики или меры расстояния [27, 32], является сложной проблемой, не имеющей однозначного решения: выбор квадрата евклидова расстояния дает более компактные кластеры, выбор прямоугольной метрики при этом делает модель менее чувствительной к выбросам. На практике приходится использовать сложные специальные способы нормирования данных [27, 78].

Сама по себе задача выделения выбросов также весьма сложна. В [27] при применении модели k-средних она решается вводом специального критерия, пороговое значение которого достаточно трудно определить экспериментально.

Как и классический алгоритм k-средних, EM-алгоритм является простой двухшаговой процедурой: шаг определения параметров кластеров и шаг разбиения на кластеры, что позволяет в перспективе использовать те же методы повышения точности и стабильности получаемых решений, которые успешно применяются при решении задач автоматической группировки электрорадиоизделий с моделями k-средних, k-медоид и k-медиан [27]. Методы, основанные на указанных моделях, замечательно себя зарекомендовали на достаточно больших выборках [27] – от нескольких сотен экземпляров электрорадиоизделий. В то же время, некоторые наиболее дорогие виды электрорадиоизделий, несущие в космических аппаратах наибольшую функциональную нагрузку, поступают партиями от нескольких штук, что делает применение EM-алгоритма и алгоритмов на его основе весьма перспективным применительно к нашей задаче.

Известные реализации EM-алгоритма, например, в среде моделирования R [67], включают пакеты EMCluster [53] и пакет mclust [54]. В данных пакетах реализованы версии EM-алгоритма и различные способы инициализации для нормального распределения смесей. Предусмотрена возможность визуализации результатов работы.

Мы использовали реализацию EM-алгоритма в среде моделирования R для автоматической группировки партий ЭРИ от 50 до 620 штук (рис.2.1). Кроме определения принадлежности каждого ЭРИ (точки на диаграмме) к кластеру

(выделены цветом), алгоритм дает таблицу с вероятностными характеристиками принадлежности к кластеру и показывает форму кластера, визуализируя возможные корреляции параметров ЭРИ.

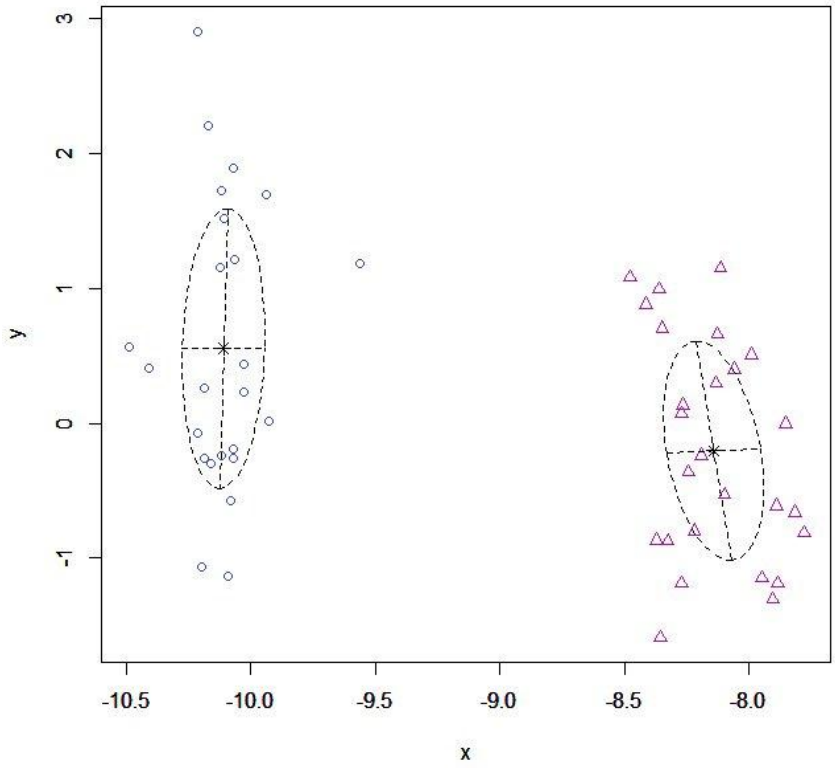


Рисунок 2.1 – Срез (показаны 2 параметра из 32) результата автоматической группировки усилителей 140УД25АС1ВК.

Использование ЕМ-алгоритма и алгоритмов на его основе открывает новые перспективы в решении задачи автоматической группировки электрорадиоизделий по производственным партиям, в частности, при небольшом объеме входных данных. При этом требуется решить проблему повышения устойчивости результата алгоритма.

Экспериментально установлено, что основной ЕМ-алгоритм обладает сильной неустойчивостью по начальным данным. Например, в случае четырехкомпонентной смеси нормальных законов при объеме выборки 200-300 наблюдений замена лишь одного наблюдения другим может кардинально поменять итоговые оценки, полученные с помощью ЕМ-алгоритма 152]. Для борьбы с такого рода неустойчивостью были предложены [39] медианные модификации основного ЕМ-алгоритма. Основной идеей этих модификаций состоит в том, чтобы заменить

наиболее «неустойчивые» этапы (моментные оценки наибольшего правдоподобия на М-шаге) на более устойчивые робастные оценки медианного типа. Эти оценки являются оптимальными в смысле среднего абсолютного отклонения. В этих модификациях в качестве оценки параметров α_j на*(m+i)*-й итерации предлагается использовать медиану случайной величины, полученную при переупорядочивании значений по неубыванию. Таким образом, медианная модификация более устойчива по наличию «засоряющих» наблюдений, т.е. относящихся к другим компонентам смеси. В [39] предлагаются два варианта медианной модификации основного ЕМ-алгоритма.

В основе стохастической (или случайной) модификации ЕМ-алгоритма лежит простой способ устранения недостатка основного алгоритма «скатываться» к ближайшему локальному максимуму функции правдоподобия. Этот способ состоит в том, чтобы, не ограничиваясь единственной траекторией базового алгоритма (находящей ближайшее локальное решение), выполнить несколько траекторий алгоритма, используя несколько случайно заданных начальных решений с последующим выбором решения с наибольшим значением функции правдоподобия. Такой подход не отвечает на вопрос о выборе оптимального механизма перехода между начальными решениями.

Другой способ состоит в целенаправленном случайном «встряхивании» наблюдений каждой итерации. Этот способ предложен в основе SEM-алгоритма (Stochastic EM) – стохастического или случайного ЕМ-алгоритма [106].

SEM-алгоритм работает относительно быстро по сравнению с другими методами. Его результат работы практически не зависит от начального приближения, он позволяет избегать выхода на неустойчивые локальные максимумы функции правдоподобия. Это достигается благодаря постоянному «встряхиванию» выборки.

Другая модификация классического ЕМ-алгоритма работает по принципу четкой классификации данных выборки (CEM-алгоритм Classification EM algorithm). Данная модификация алгоритма совпадает с SEM, за исключением того, что на каждом шаге вводится детерминированное правило, относящее данные

только к одному из кластеров, для которого вычислили максимальную апостериорную вероятность.

Модификации EM-алгоритма (SEM, CEM) улучшают работу базового алгоритма, но при этом все они не лишены своего главного недостатка –это варианты локального поиска, не дающие удовлетворительных результатов для поиска глобального решения. При этом данные алгоритмы не показывают стабильного результата для многомерных данных.

2.2 Жадные эвристические процедуры для задач разделения смеси

Идея настоящей работы состоит в применении подходов метода жадных эвристик к задаче автоматической группировки на основе разделения смеси распределений. Основной принцип работы жадных эвристических алгоритмов (далее по тексту - *жадные эвристики*) показан на рис. 2.2.

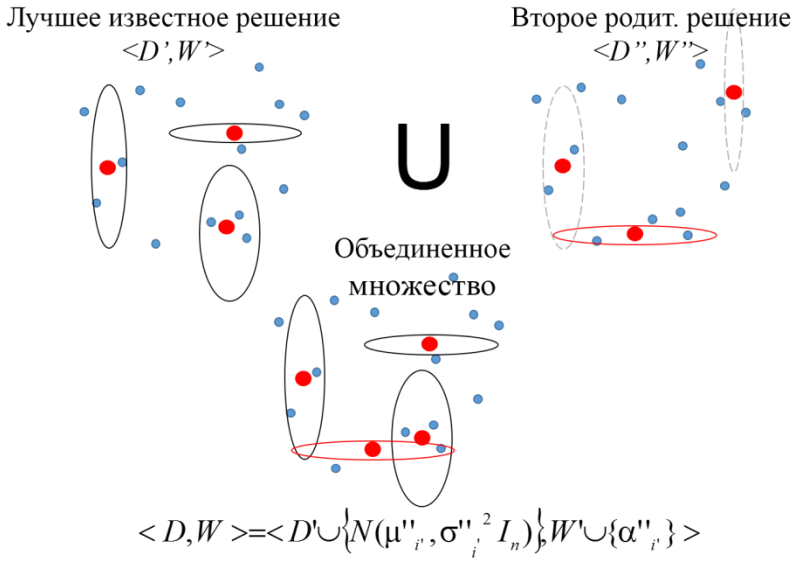


Рисунок 2.2 – Принцип объединения множеств родительских решений для задачи разделения смеси распределений.

Жадная агломеративная эвристика для задачи разделения смеси состоит из двух шагов:

Дано: Имеются два известных (родительских) решения задачи.

Решения представлены парами множеств <*D, W*>. Множество *D* это множество распределений в смеси. Каждое распределение задано параметрами. Второе множество *W* – это соответствующие распределениям априорные вероятности.

Шаг 1. Берутся два родительских решения (первое из которых, например, является лучшим из известных). Множества этих решений объединяются. Получается промежуточное решение. Оно недопустимое – избыточной мощности, т.е. с избыточным числом распределений.

Шаг 2. В жадной эвристике происходит последовательное уменьшение числа распределений. Каждый раз отсекается то распределение, удаление которого даёт наименьшее ухудшение значения целевой функции.

Ниже представлен алгоритм работы базовой жадной эвристики:

Алгоритм 2.1 Базовая жадная агломеративная эвристика для разделения гауссовых распределений (на примере сферических гауссовых распределений).

Дано: начальное число гауссовых распределений (кластеров) *K*, требуемое число кластеров *k*. При этом *K*>*k*.

1. Выбрать начальное решение с *K* кластерами, т.е. выбрать случайным образом начальные параметры пары множеств распределений и их весовых коэффициентов $< D,W >=< \{N(\mu_i^{\langle 0\rangle}, \sigma_i^{\langle 0\rangle 2} I_n)\}, \{\alpha_i^{\langle 0\rangle}\}, i = \overline{1,K} >$.

2. Выполнить Алгоритм 1.1 (EM-алгоритм), 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 $< D', W' >=< D \setminus \{N(\mu_{i'}^{\langle 0\rangle}, \sigma_{i'}^{\langle 0\rangle 2} I_n)\}, \quad W \setminus \{\alpha_{i'}^{\langle 0\rangle}\} >$.

4.2. Запустить Алгоритм 1.1 с начальными значениями параметров распределений, представленных усеченным <*D', W'*>. При этом Алгоритм 1.1 ограничивается одной итерацией. Для полученного EM-алгоритмом решения

рассчитать целевую функцию L согласно (1.2), сохранить ее значение в переменной $L'_{i'}$.

4.3. Следующая итерация цикла 4.

5. Найти индекс $i'' = arg \max\limits_{i'=\overline{1,k}} L_{i'}$.

6. Получить пару усеченных $< D'', W'' >=< D \setminus \{N(\mu_{i''}^{\langle 0 \rangle}, \sigma_{i''}^{\langle 0 \rangle 2} I_n)\}, \quad W \setminus \{\alpha_{i''}^{\langle 0 \rangle}\} >$. Запустить для этой пары усеченных множеств Алгоритм 1.1, затем перейти к шагу 3.

В ходе работы нами были предложены новые эвристические процедуры.

Ниже приводится описание жадных эвристических процедур и результаты вычислительных экспериментов.

Алгоритм 2.2. Жадная процедура с частичным объединением №1.

Дано: пары множеств распределений

$$< D', W' >=< \{N(\mu'^{\langle 0 \rangle}_i, \sigma'^{\langle 0 \rangle 2}_i I_n)\}, \{\alpha'^{\langle 0 \rangle}_i\}, i = \overline{1,K} >$$

и

$$< D'', W'' >=< \{N(\mu''^{\langle 0 \rangle}_i, \sigma''^{\langle 0 \rangle 2}_i I_n)\}, \{\alpha''^{\langle 0 \rangle}_i\}, i = \overline{1,K} >.$$

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

1.1. Объединить поэлементно множества в парах $<D',W'>$ и $<D'',W''>$: $< D, W >=< D' \cup \{N(\mu''^{\langle 0 \rangle}_{i'}, \sigma''^{\langle 0 \rangle 2}_{i'} I_n)\}, W' \cup \{\alpha'^{\langle 0 \rangle}_{i''}\} >$.

1.2. Запустить базовую жадную эвристику (Алгоритм 2.1) с этими парами объединенных множеств $<D,W>$ в качестве начального решения. Полученный результат (пару полученных множеств, а также значение целевой функции) сохранить.

3. 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, представленный парой $<D', W'>$ поочередно каждым из элементов второго «родительского» решения, представленного

соответствующей парой множеств <*D’’, W’’*>, затем к этой паре объединенных множеств применяет базовую жадную эвристику (Алгоритм 2.1) и фиксирует лучший из полученных результатов.

Более простой, но более требовательный в плане вычислительных ресурсов вариант аналогичного алгоритма представлен ниже.

Алгоритм 2.3. Жадная процедура с полным объединением родительских решений.

Дано: см. Алгоритм 2.2

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

$< D,W >=< D' \cup D'' ,\quad W' \cup W'' >$.

2. Запустить Алгоритм 2.1 с этими объединенными множествами в качестве начального решения.

Как видно, здесь выполняется полное объединение двух «родительских» вариантов решения задачи разделения смесей распределений, представленных парами множеств <*D’, W’*> и <*D’’, W’’*> соответственно. Наконец, возможен промежуточный вариант, в котором множества объединяются частично, при этом первое множество берется полностью, а из второго множества выбирается случайным образом случайное число элементов. Подобный подход хорошо зарекомендовал себя при решении задач автоматической группировки с применением моделей k-средних, k-медоид, k-медиан [29, 152].

Алгоритм 2.4. Жадная процедура с частичным объединением №2.

Дано: см. Алгоритм 2.2

1. Выбрать случайное $r \in \overline{\{2,k-1\}}$ с равной вероятностью.

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*).

2.2. Объединить множества $< D,W >=< D' \cup D'',\ W' \cup W'' >$.

2.3. Запустить Алгоритм 2.1 с этими объединенными множествами в качестве начального решения.

3. Возвратить в качестве результата лучшее (по значению целевой функции) из решений, полученных на шаге 2.3.

Первые же вычислительные эксперименты показали крайнюю неэффективность Алгоритма 2.4 в сравнении с Алгоритмом 2.3 для всех исследованных задач. В то же время, эффективность значительно повышается, если ограничить число элементов, добавляемых из решения *<D'', W''>* следующим образом:

Шаг 1 Алгоритма 2.4: Выбрать случайное $r' \in [0;1)$. Присвоить r=[(k/2-2) r'2]+2. Здесь [.] – целая часть числа.

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 базовой жадной агломеративной эвристики (Алгоритма 2.1).

2.3 Новый алгоритм поиска в чередующихся окрестностях для задачи разделения смеси распределений

Метод жадных эвристик [27] предусматривает в качестве одного из вариантов организации локального поиска применение алгоритмов поиска в чередующихся окрестностях (VNS-алгоритмы –Variable Neighborhood Search) [131, 171, 134, 133]. В [42] приводится обзор современных методов локального поиска, основанных на идее чередующихся окрестностей (в частности, алгоритма поиска с

чередующимися окрестностями). Показаны способы комбинирования этих методов с другими метаэвристиками. Приводятся примеры удачного применения данных методов при поиске дискретных структур, существование которых представлялось проблематичным. На примере задачи коммивояжера демонстрируются разные способы построения окрестностей и их свойства.

Суть алгоритма поиска с чередующимися окрестностями заключается в том, что для некоторого промежуточного решения X есть конечное множество окрестностей S_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 классической задачей непрерывной оптимизации. В качестве метода локального поиска вполне успешно может использоваться EM-алгоритм и его модификации (CEM, SEM, медианный EM [39, 105]). С одной стороны, алгоритм глобального поиска должен периодически «выбивать» промежуточное решение задачи из «области притяжения» локального оптимума, а с другой стороны, решения, образованные из элементов различных локально-оптимальных решений, с большей вероятностью оказываются ближе к глобальному оптимуму, чем случайно выбранные решения [79]. Таким образом, перспективным представляется поиск в окрестности локального оптимума, образованной заменой отдельных элементов локально-оптимального решения элементами других локально-оптимальных решений, что, собственно и выполняют различные варианты Алгоритма 2.4, промежуточные решения, которого представлены парами множеств распределений и их весовых коэффициентов, каждая из которых является результатом работы EM-алгоритма, то есть локальным максимумом. Другими словами, в данной работе предлагается использовать VNS-алгоритм в качестве расширенного локального поиска.

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 2.2, 2.3, 2.4. Таким образом, эти алгоритмы осуществляют поиск в некоторых окрестностях решения <*D', W'*>, а второе решение <*D'', W''*> неявным образом задает параметры этой окрестности. Так, Алгоритм 2.2 осуществляет поиск в окрестности $S(<D', W'>)=\{ greedy(<D',W'> \cup \{<D''_i, W''_i>\}), i=1,|D''|\}$. Здесь *greedy( )* – результат применения Алгоритма 2.1 Соответственно, Алгоритмы 2.3 и 2.4 осуществляют поиск в более широких окрестностях.

Таким образом, общую схему алгоритма поиска в чередующихся окрестностях для решения задач разделения смеси распределений можно описать следующим образом:

Алгоритм 2.5. Поиск в чередующихся окрестностях для разделения смеси распределений.

1. Запустить EM-алгоритм из случайного начального решения, получить решение <*D, W*>.

2. Установить $s=s_{start}$ (номер окрестности поиска)

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

4. Запустить EM-алгоритм из случайного начального решения, получить решение <*D', W'*>.

5. В зависимости от значения *s* (возможны значения 1, 2 или 3), запустить Алгоритм 2.2, 2.4 или 2.3 с начальными решениями <*D, W*> и <*D', W'*>. Таким образом окрестность определяется процедурой включения распределений из второго известного решения (указанные Алгоритмы 2.2, 2.4 или 2.3) и параметром окрестности – собственно вторым известным решением. Поиск осуществляется в этой определенной окрестности.

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 $i=0$, $j=j+1$. Осуществить переход к новой окрестности: $s=s+1$; 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.

Важными являются значения двух параметров: i_{max} – число безрезультатных поисков в окрестности и j_{max} – число безрезультатных переключений окрестностей. Для j_{max} вполне достаточно значения $j_{max}=2$. Мы использовали значение $i_{max}=2k$.

Также важным является параметр s_{start}, задающий номер окрестности, с которой начинается поиск. Данный параметр особенно важен. Мы провели вычислительные эксперименты со всеми возможными его значениями. В зависимости от этого значения алгоритмы обозначены ниже соответственно VNS1, VNS2, VNS3. Старт алгоритма поиска может начинаться с разных окрестностей.

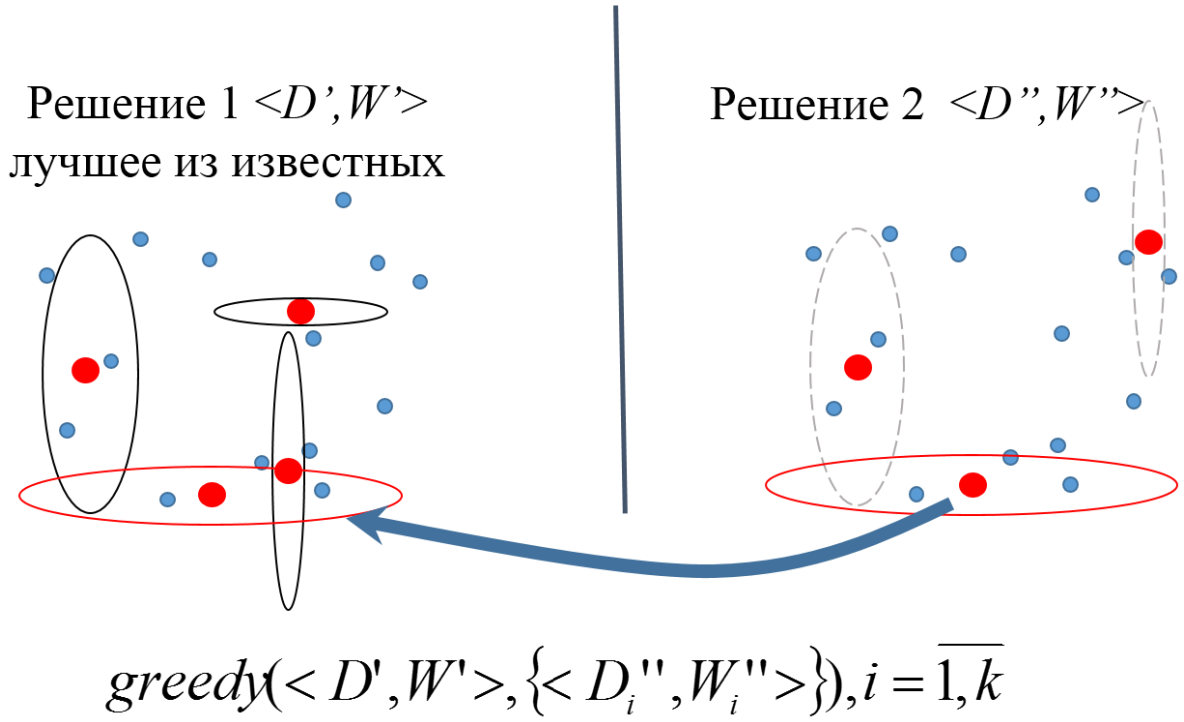


Рисунок 2.3 – Один из принципов формирования окрестности решений для задачи разделения смеси распределений (VNS1).

На рис. 2.3 показана графическая интерпретация принципа формирования

окрестностей решений для VNS1.

Вычислительные эксперименты показывают высокую конкурентоспособность различных версий Алгоритма 2.5 для определенных классов задач.

Для тестирования Алгоритма 2.5 в трех различных его модификациях мы использовали как результаты тестирования электрорадиоизделий, полученные ОАО «ИТЦ НПО-ПМ» (эти данные нормированы специальным образом – по границам дрейфа), так и классические наборы данных из репозитория UCI (www.cs.uci.edu/mlearn/mlrepository.html).

Результаты (Таблица 2.1) показывают, что новый алгоритм поиска в чередующихся окрестностях для разделения смеси распределений не имеет преимущества перед мультистартом классического EM-алгоритма при очень малом числе кластеров (менее 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). При этом невозможно отдать однозначное предпочтение одной из версий нового алгоритма.

Для большинства наборов данных было выполнено по 30 попыток запуска каждого из алгоритмов, для наборов данных KDDCUP04Bio и Europe – по 10 попыток. 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.

Таблица 2.1 – Сравнительные результатыVNS и модификаций EM-алгоритма

Набор данных, число вектор., размерн.	Число кластеров *k*, тип распр., время	Алгоритм	Средний результат (целевая функция)	Среднеквадратичное откл. результатов
Ionosphere, *N*=351, *d*=35, нормирован.	10, сфер., 30 сек.	VNS1 VNS2 VNS3 EM CEM SEM	-847,463* -849,352 -878,153 -871,405 -893,442 -879,952	23,972 28,073 41,875 15,786 7,879 15,192
KDDCUP04 Bio, *N*=145751, *d*=74, нормир.	30, сфер., 3 часа	VNS1 VNS2 VNS3 EM CEM SEM	-12511968,2* -12511984,9 -12512654,0 -12513229,7 -12512717,9 -12514336,5	105,255 278,035 139,040 1255,541 343,026 683,607
Europe (UCI), *N*=169308, *d*=2.	40, сфер, 1.5 часа	VNS1 VNS2 VNS3 EM CEM SEM	-3625694,1 -3625691,7 -3625748,7 -3625957,3 -3625779,0 -3409382,2*	20,148 14,769 15,402 49,561 25,064 29,064
Тесты ИС 140УД17АВК, *N*=51, *d*=46	5, сфер., 5 сек.	VNS1 VNS2 VNS3 EM	3673,671* 3543,974 3591,539 3598,16	44,043 144,691 139,988 32,16

Примечание: * - лучший результат.

2.4 Генетический алгоритм метода жадных эвристик с вещественным алфавитом для задач разделения смеси распределений

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 $D_l = \{N(\mu_{l,i}, \sigma_{l,i}{}^2 I_n), l = \overline{1,k}\}, l = \overline{1, N_{POP}}$ и $W_l = \{\alpha_{l,i} = 1/k, i = \overline{1,k}\}$, где N_{POP} – размер популяции алгоритма, т.е. количество «особей» - промежуточных решений, которым он оперирует.

Алгоритм 2.6. Генетический алгоритм с жадной эвристикой (дано общее описание трех вариантов, названных GA-FULL, GA-ONE, GA-RAND). Дано: Начальный размер популяции N_{POP}.

1. Сгенерировать случайным образом NPOP начальных решений, представленных множествами распределений $D_l = \{N(\mu_{l,i}, \sigma_{l,i}{}^2 I_n), l = \overline{1,k}\}, l = \overline{1, N_{POP}}$ и соответствующими множествами весовых коэффициентов $W_l = \{\alpha_{l,i} = 1/k, i = \overline{1,k}\}$. 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 1.1, полученные значения целевой функции сохраняются в переменных $f_1, \ldots, f_{N_{POP}}$. Присвоить N_{iter}=0;

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

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

2.4 (в варианте алгоритма GA-FULL – генетический алгоритм с жадной эвристикой с полным объединением), Алгоритм 2.3 (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 турнирное замещение: 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.

Для тестирования Алгоритма 2.6 в трех различных его модификациях и сравнения результатов с результатами классического EM-алгоритма, запускаемого в режиме мультистарта, мы использовали как результаты тестирования электрорадиоизделий, полученные ОАО «ИТЦ НПО-ПМ» (эти данные нормированы специальным образом – по границам дрейфа [79]), так и классические наборы данных из репозитория UCI (www.cs.uci.edu/mlearn/mlrepository.html).

Как и в случае алгоритмов с чередующимися окрестностями, новые генетические алгоритмы неэффективны для малых задач и задач с небольшим числом кластеров. Результаты (см. таблицу 2.2) показывают, что разработанные генетические алгоритмы не имеют преимущества перед мультистартом классического EM-алгоритма при очень малом числе кластеров (5 кластеров). В то же время, с ростом числа кластеров и с ростом объема выборки сравнительная эффективность повышается, и для больших задач новые алгоритмы имеют безусловное преимущество. Также новые алгоритмы оказались неэффективны для разбиения наборов булевых данных (см.набор данных Chess). При этом невозможно отдать однозначное предпочтение одной из версий нового генетического алгоритма. В этой связи логичным было бы отдать предпочтение версии, комбинирующей обе разработанные жадные эвристические процедуры в случайном порядке (вариант алгоритма GA-RAND), который дает в большинстве

случаев лучшие результаты либо близкие к лучшим.

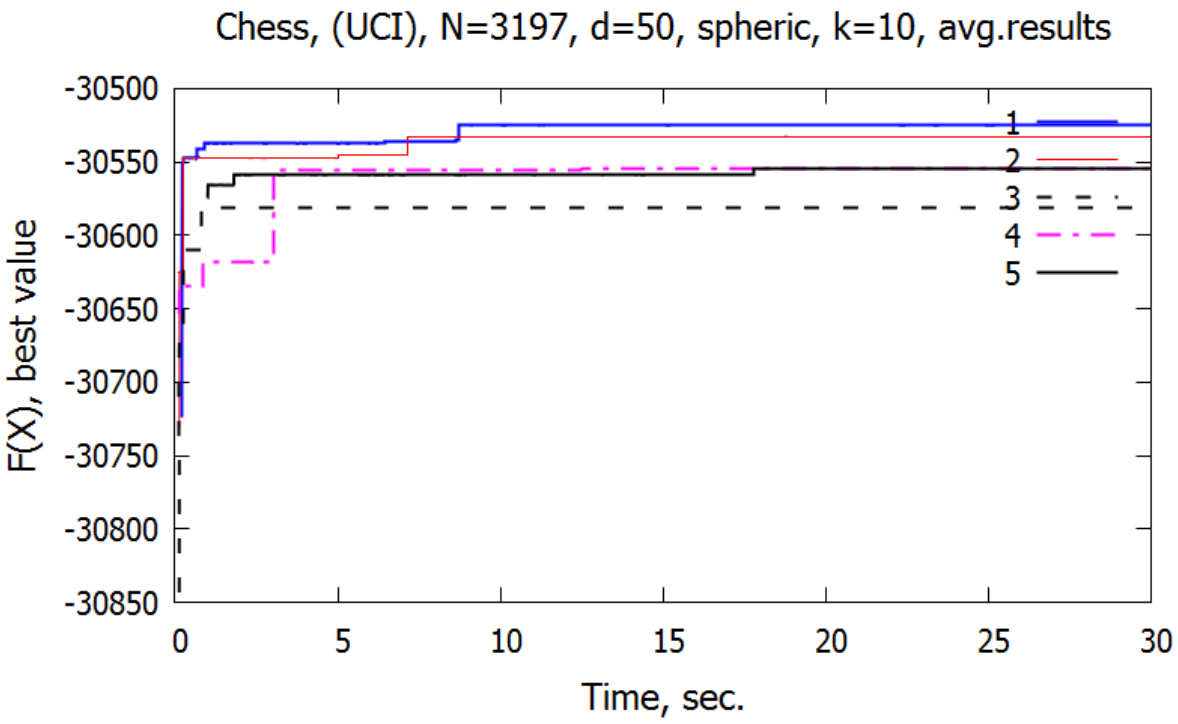
Для большинства наборов данных было выполнено по 30 попыток запуска каждого из алгоритмов, для наборов данных KDDCUP04Bio и Europe – по 10 попыток. Фиксировались только лучшие результаты, достигнутые каждым из алгоритмов в каждой попытке, затем результаты тридцати или десяти попыток были усреднены.

Таблица 2.2 – Сравнительные результаты генетических алгоритмов и модификаций EM-алгоритма

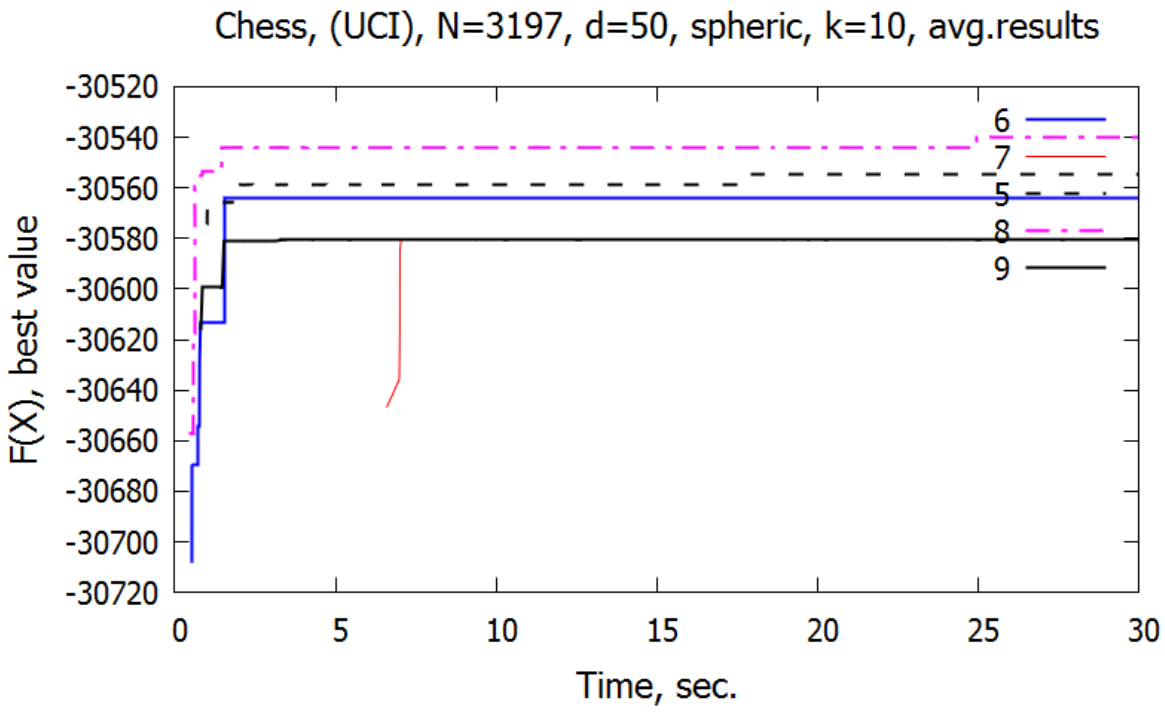
Набор данных, число векторов данных, размерность	Число кластеров *k*, тип распр-й, время	Алго-ритм	Средний результат (целеваяфунк-ция)
Chess (UCI), *N*=3197, *d*=50, булевы	10, сфер., 3мин.	EM FULL ONE RAND	-30525,03* -30581,09 -30532,44 -30554,67
Тесты ИС 140УД17АВК, *N*=51, *d*=46	2, сфер., 5 сек.,	EM FULL ONE RAND	1052942,6* 1052942,0 1052847,0 1052941,2
Тесты ИС 1526ЛЕ2, *N*=3987, *d*=206	5, некоррелир., 3 мин.	EM FULL ONE RAND	340951,0 366487,3 350292,0 443491,0*
KDDCUP04 Bio, *N*=145751, *d*=74, нормир.	30, сфер., 3 часа	EM FULL ONE RAND	-12513230 -12512517* -12512818 -12512811
Europe (UCI), *N*=169308, *d*=2.	40, сфер, 1.5 часа	EM FULL ONE RAND	-3623957,4 -3625740,6* -3625878,2 -3625816,2
Ionosphere, *N*=351, *d*=35, нормирован.	10, сферические, 30 сек.	EM FULL ONE RAND	-871,405 -960,318 -824,846* -832,828

Примечание: * - лучший результат.

Ниже на рис. 2.4 – 2.13 приведены графики сходимости алгоритмов. Сравнительный анализ новых алгоритмов: ГА и на основе поиска в чередующихся окрестностях (VNS) с модификациями EM-алгоритма показывает, что новые алгоритмы имеют преимущество.



А)

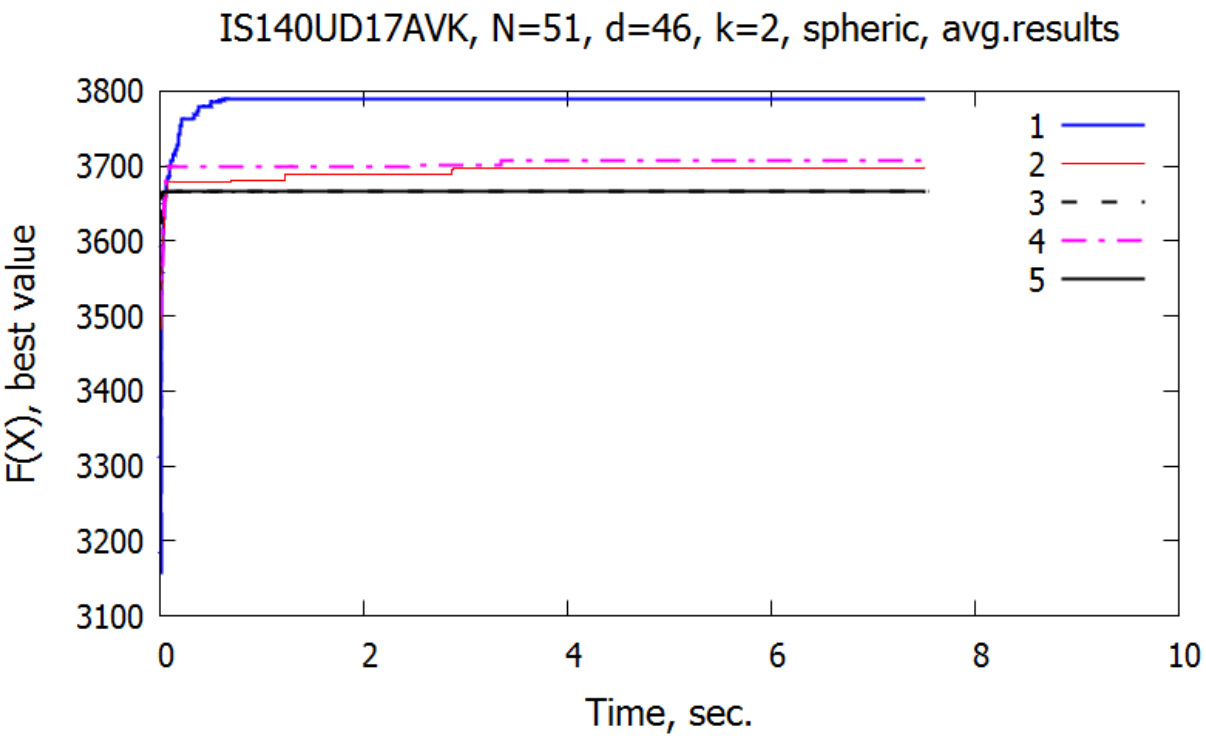


Б)

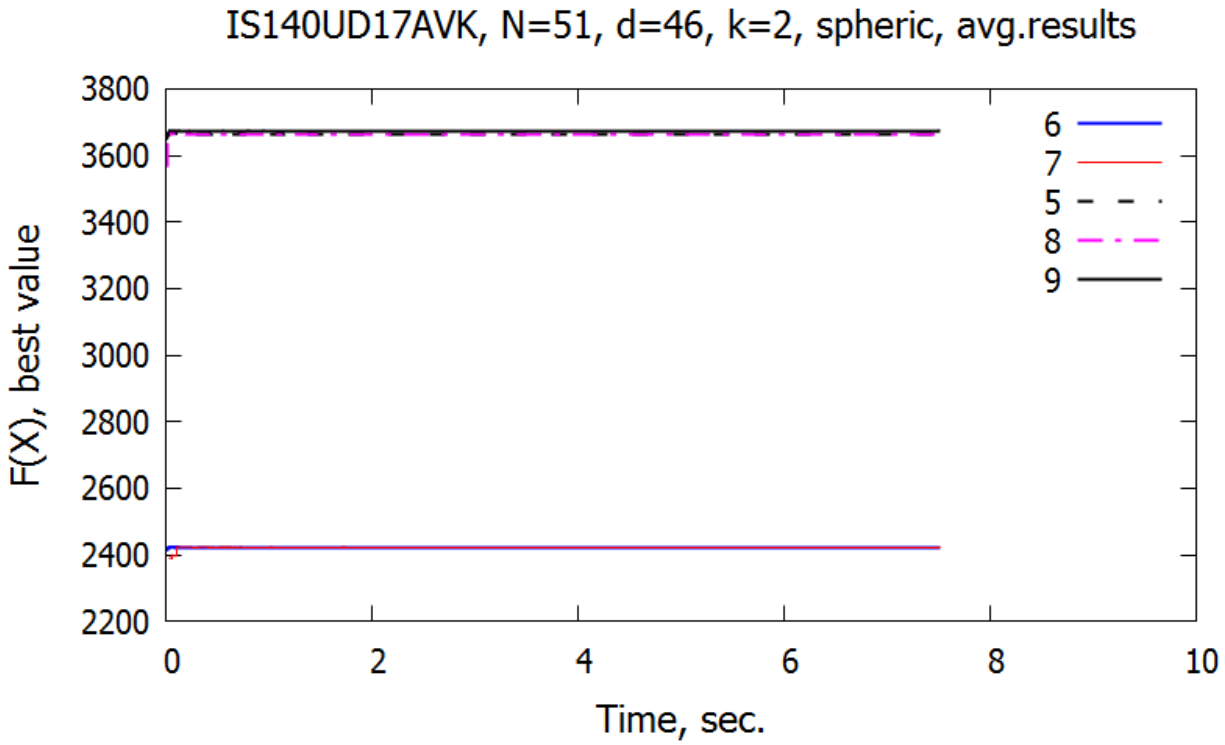
Рисунок 2.4 –Сравнение новых алгоритмов (ГА и VNS) и EM-алгоритмов (классический EMи его модификации CEMи SEM) для набора данных Chess (UCI).

По оси абсцисс – время в секундах, по оси ординат – достигнутое значение целевой функции

1 – EM-алгоритм (EM);
2 – модификация ГА (GA-ONE);
3 – модификация ГА (GA-FULL);
4 – модификация ГА (GA-RAND);
5 – модификация VNS алгоритма(VNS-1);
6 – CEM-алгоритм (CEM);
7 – SEM-алгоритм (SEM);
5 – модификация VNS алгоритма (VNS-1);
8 – модификация VNS алгоритма (VNS-2);
9 – модификацияVNSалгоритма (VNS-3);



А)

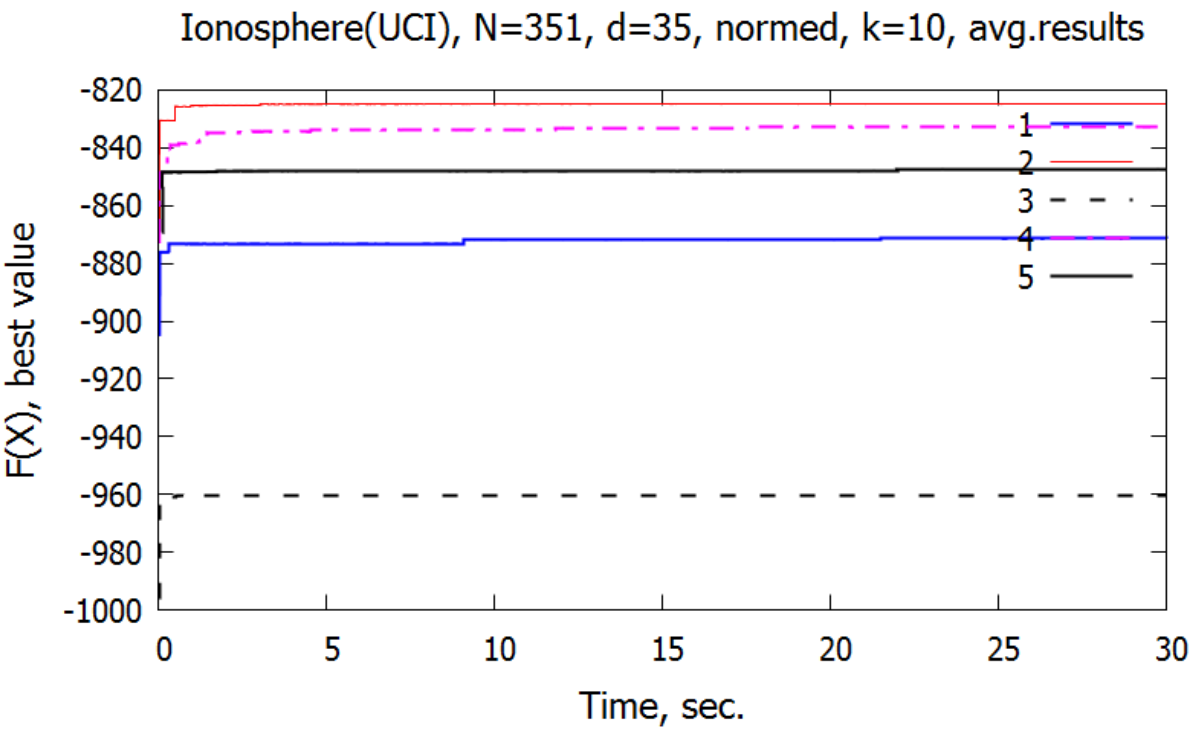


Б)

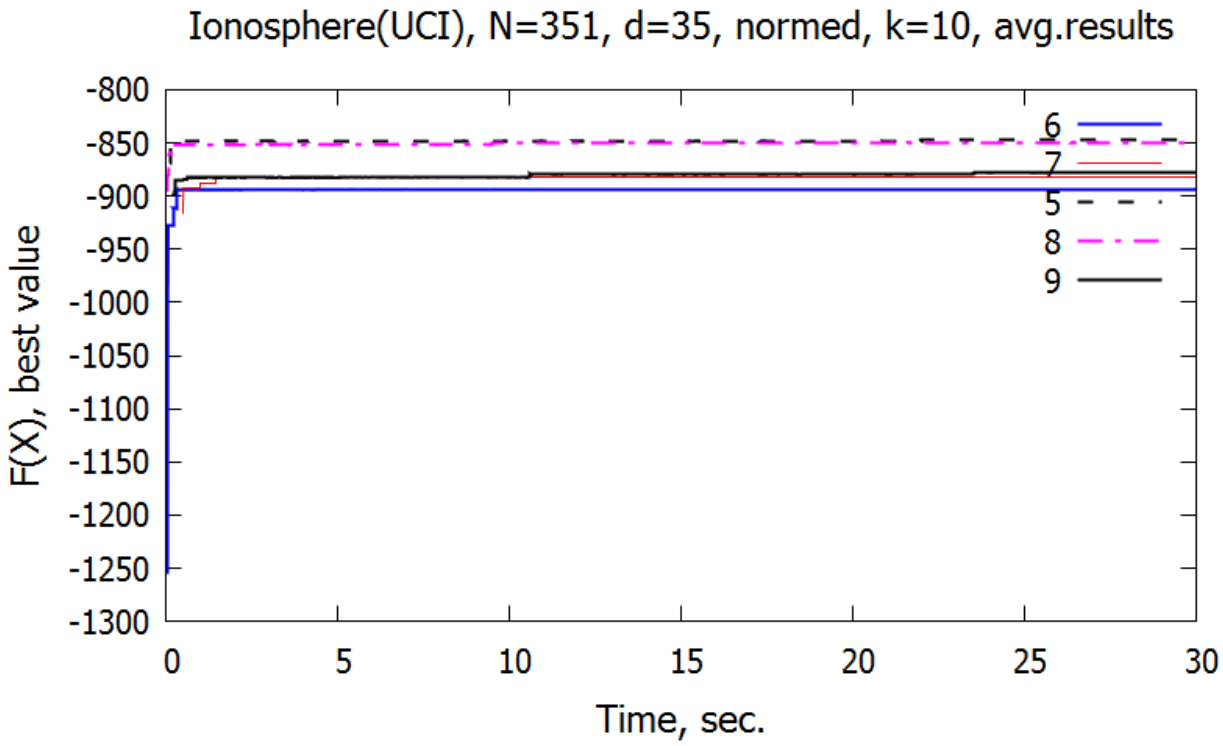
Рисунок 2.5 –Сравнение новых алгоритмов (ГА и VNS) и ЕМ-алгоритмов (классический ЕМи его модификации СЕМи SEM) для набора данных ИС140УД17АВК.

По оси абсцисс – время в секундах, по оси ординат – достигнутое значение целевой функции

1 – ЕМ-алгоритм (EM);
2 – модификация ГА (GA-ONE);
3 – модификация ГА (GA-FULL);
4 – модификация ГА (GA-RAND);
5 – модификация VNS алгоритма(VNS-1);
6 – СЕМ-алгоритм (СЕМ);
7 – SEM-алгоритм (SEM);
5 – модификация VNS алгоритма (VNS-1);
8 – модификация VNS алгоритма (VNS-2);
9 – модификацияVNSалгоритма (VNS-3);



А)

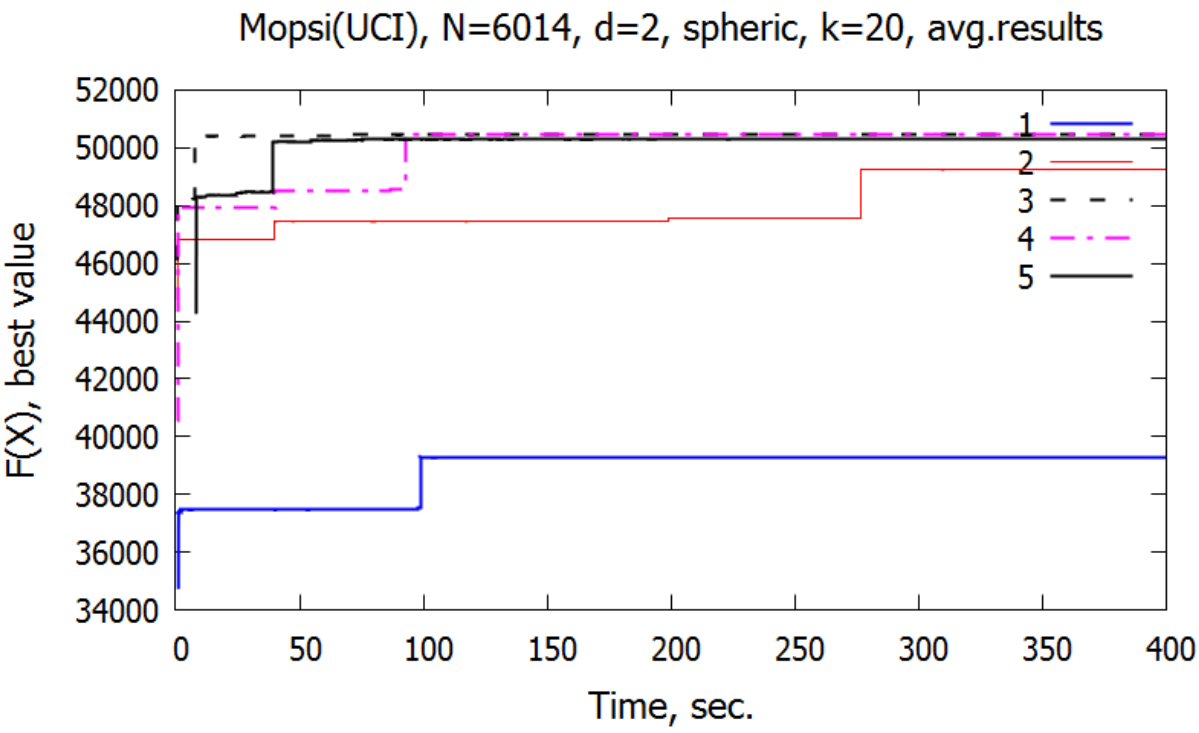


Б)

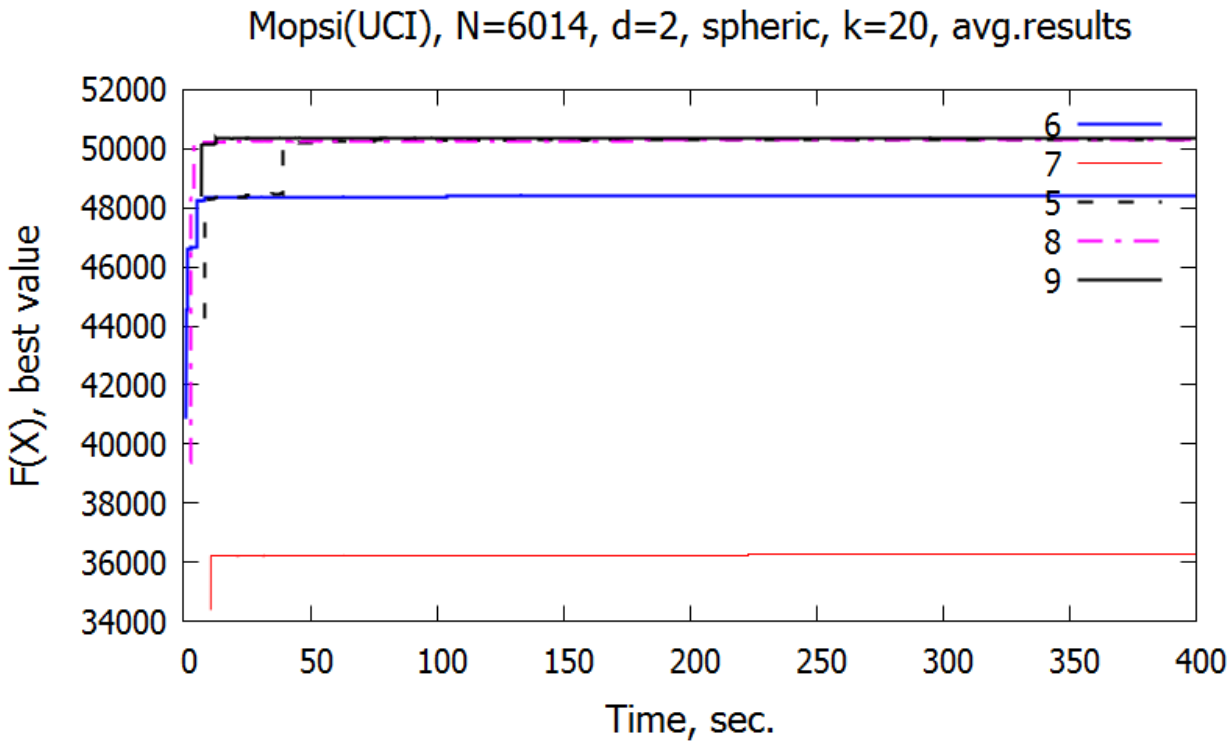
Рисунок 2.6 –Сравнение новых алгоритмов (ГА и VNS) и EM-алгоритмов (классический EMи его модификации CEMи SEM) для набора данных Ionosphere (UCI)

По оси абсцисс – время в секундах, по оси ординат – достигнутое значение целевой функции

1 – EM-алгоритм (EM);
2 – модификация ГА (GA-ONE);
3 – модификация ГА (GA-FULL);
4 – модификация ГА (GA-RAND);
5 – модификация VNS алгоритма(VNS-1);
6 – CEM-алгоритм (CEM);
7 – SEM-алгоритм (SEM);
5 – модификация VNS алгоритма (VNS-1);
8 – модификация VNS алгоритма (VNS-2);
9 – модификацияVNSалгоритма (VNS-3);



А)

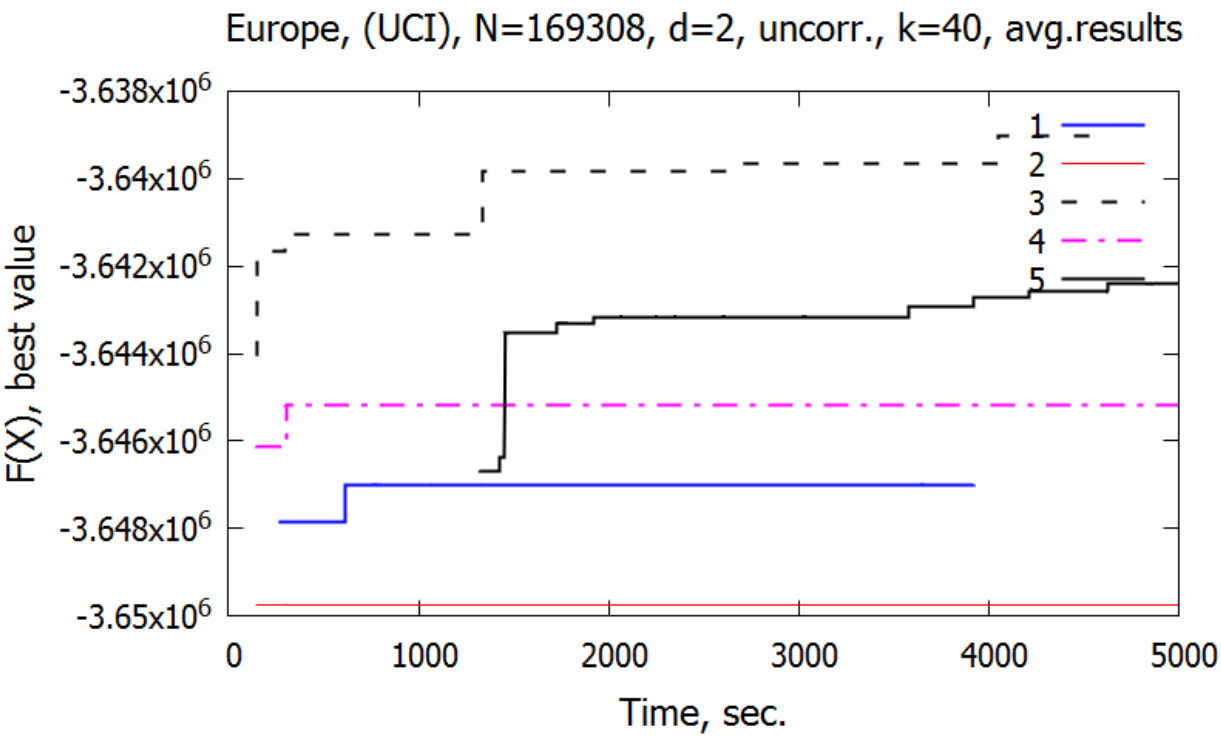


Б)

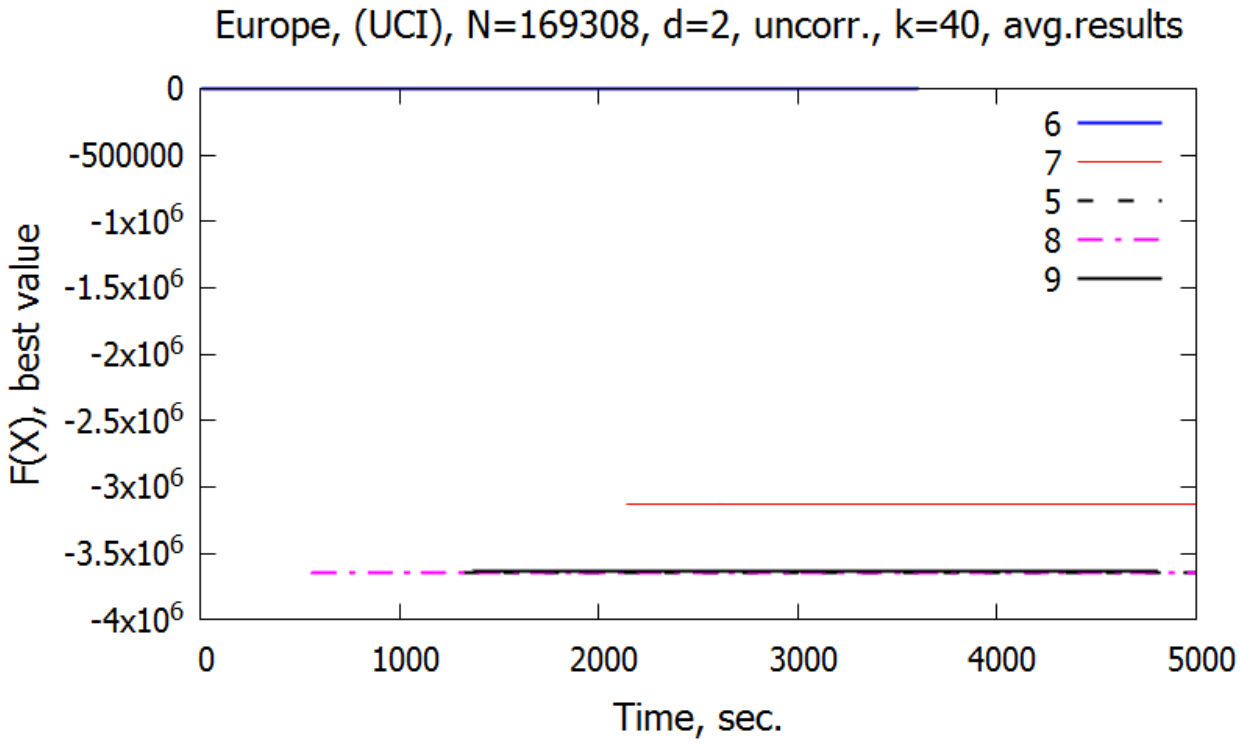
Рисунок 2.7 –Сравнение новых алгоритмов (ГА и VNS) и ЕМ-алгоритмов (классический ЕМи его модификации СЕМи SEM) для набора данных Mopsi (UCI).

По оси абсцисс – время в секундах, по оси ординат – достигнутое значение целевой функции

1 – ЕМ-алгоритм (ЕМ);
2 – модификация ГА (GA-ONE);
3 – модификация ГА (GA-FULL);
4 – модификация ГА (GA-RAND);
5 – модификация VNS алгоритма(VNS-1);
6 – СЕМ-алгоритм (СЕМ);
7 – SEM-алгоритм (SEM);
5 – модификация VNS алгоритма (VNS-1);
8 – модификация VNS алгоритма (VNS-2);
9 – модификацияVNSалгоритма (VNS-3);



А)

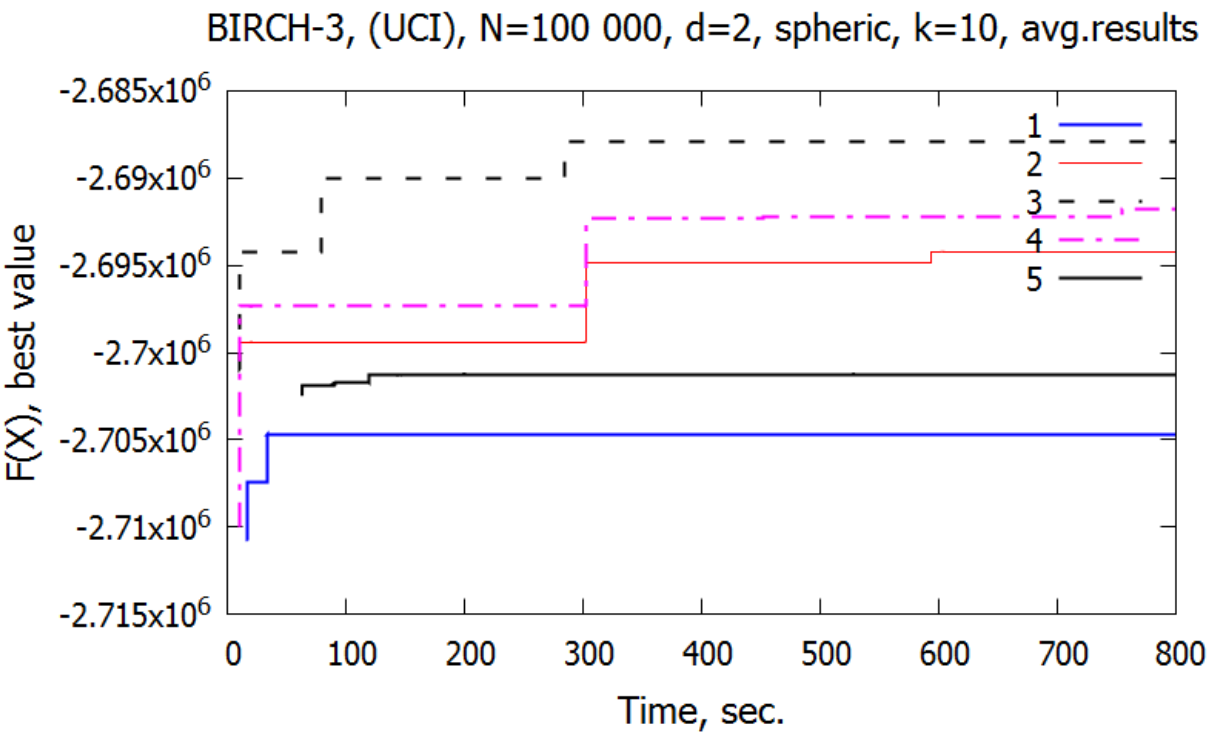


Б)

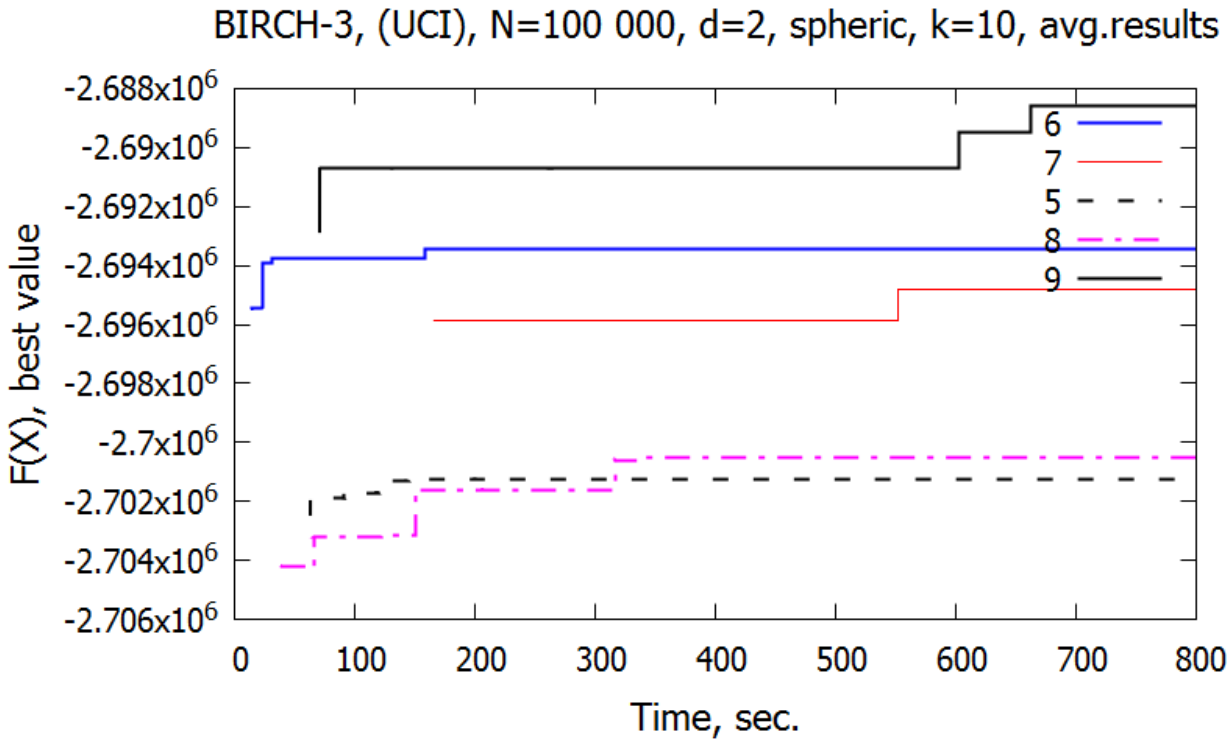
Рисунок 2.8 –Сравнение новых алгоритмов (ГА и VNS) и EM-алгоритмов (классический EMи его модификации CEMи SEM) для набора данных Europe (UCI).

о оси абсцисс – время в секундах, по оси ординат – достигнутое значение целевой функции

1 – EM-алгоритм (EM);
2 – модификация ГА (GA-ONE);
3 – модификация ГА (GA-FULL);
4 – модификация ГА (GA-RAND);
5 – модификация VNS алгоритма(VNS-1);
6 – CEM-алгоритм (CEM);
7 – SEM-алгоритм (SEM);
5 – модификация VNS алгоритма (VNS-1);
8 – модификация VNS алгоритма (VNS-2);
9 – модификацияVNSалгоритма (VNS-3);



А)

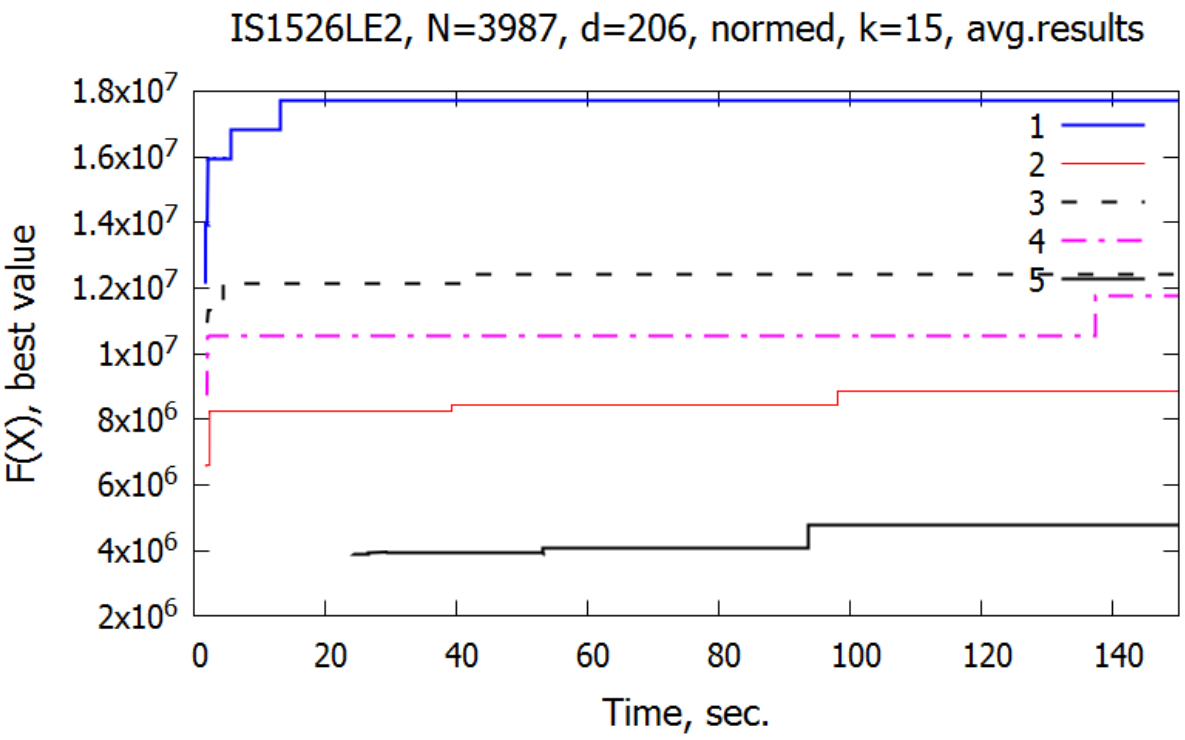


Б)

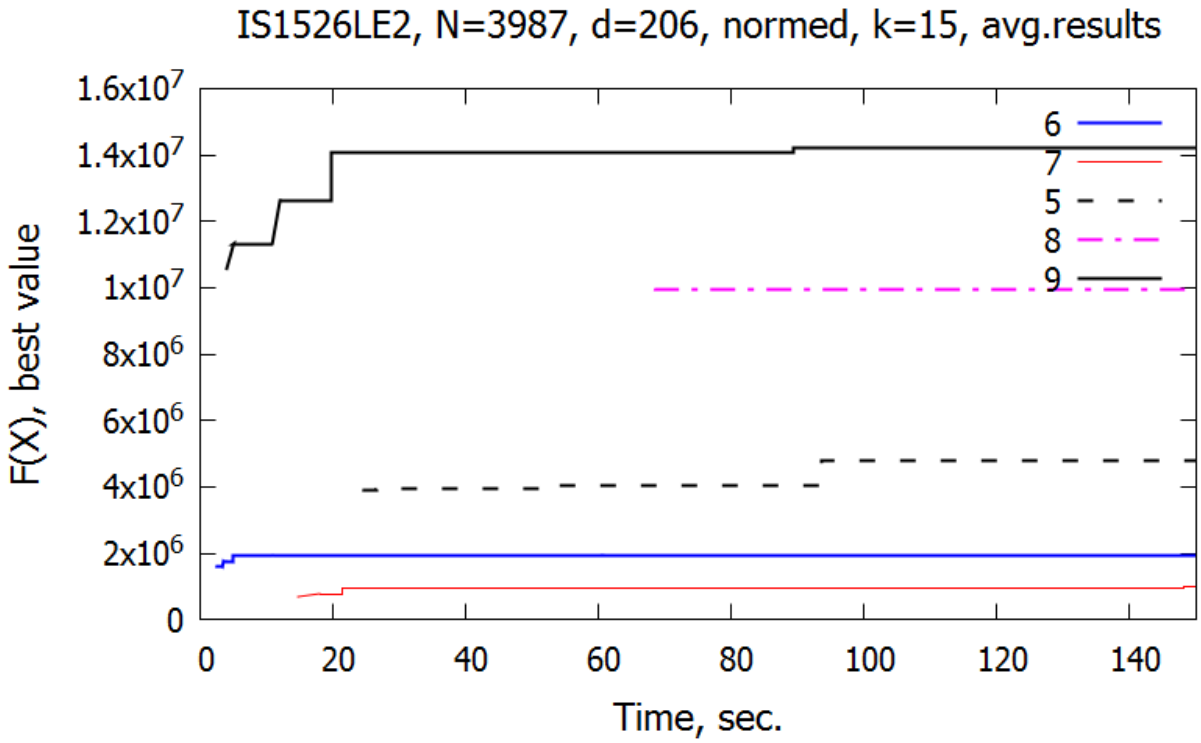
Рисунок 2.9 –Сравнение новых алгоритмов (ГА и VNS) и ЕМ-алгоритмов (классический ЕМи его модификации CEMи SEM) для набора данных BIRCH-3 (UCI).

о оси абсцисс – время в секундах, по оси ординат – достигнутое значение целевой функции

1 – ЕМ-алгоритм (EM);
2 – модификация ГА (GA-ONE);
3 – модификация ГА (GA-FULL);
4 – модификация ГА (GA-RAND);
5 – модификация VNS алгоритма(VNS-1);
6 – CEM-алгоритм (CEM);
7 – SEM-алгоритм (SEM);
5 – модификация VNS алгоритма (VNS-1);
8 – модификация VNS алгоритма (VNS-2);
9 – модификацияVNSалгоритма (VNS-3);



А)

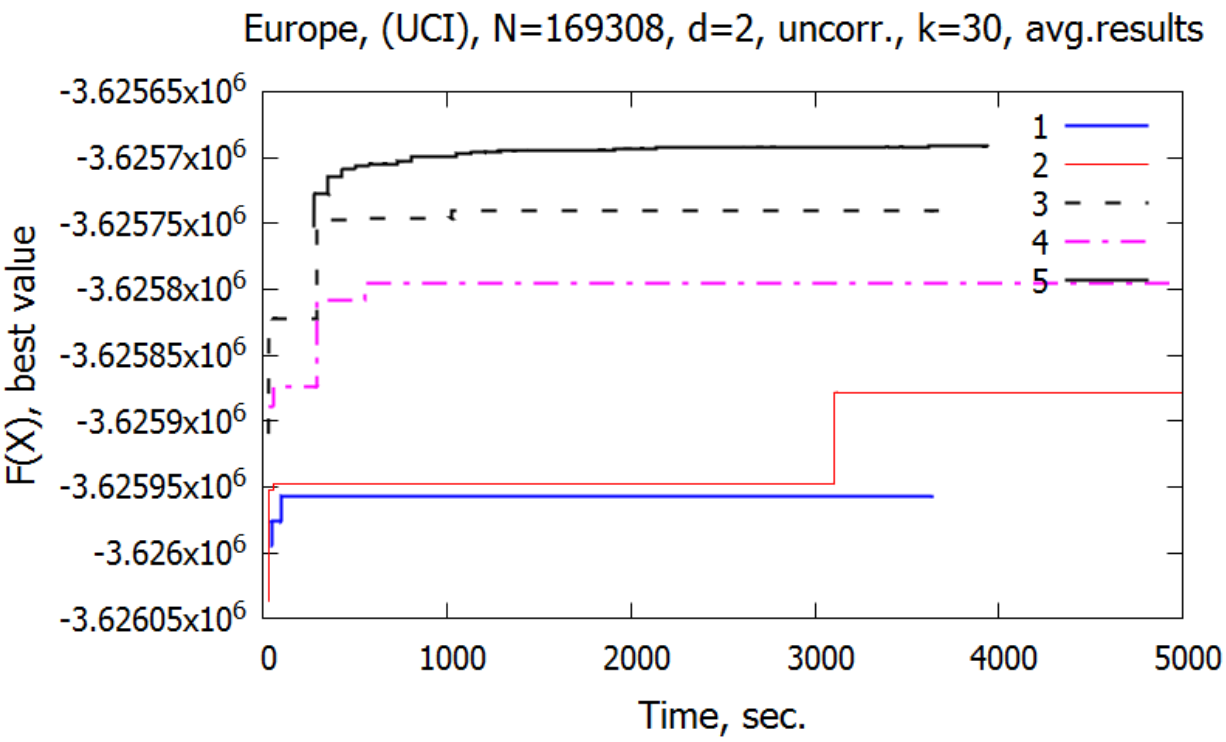


Б)

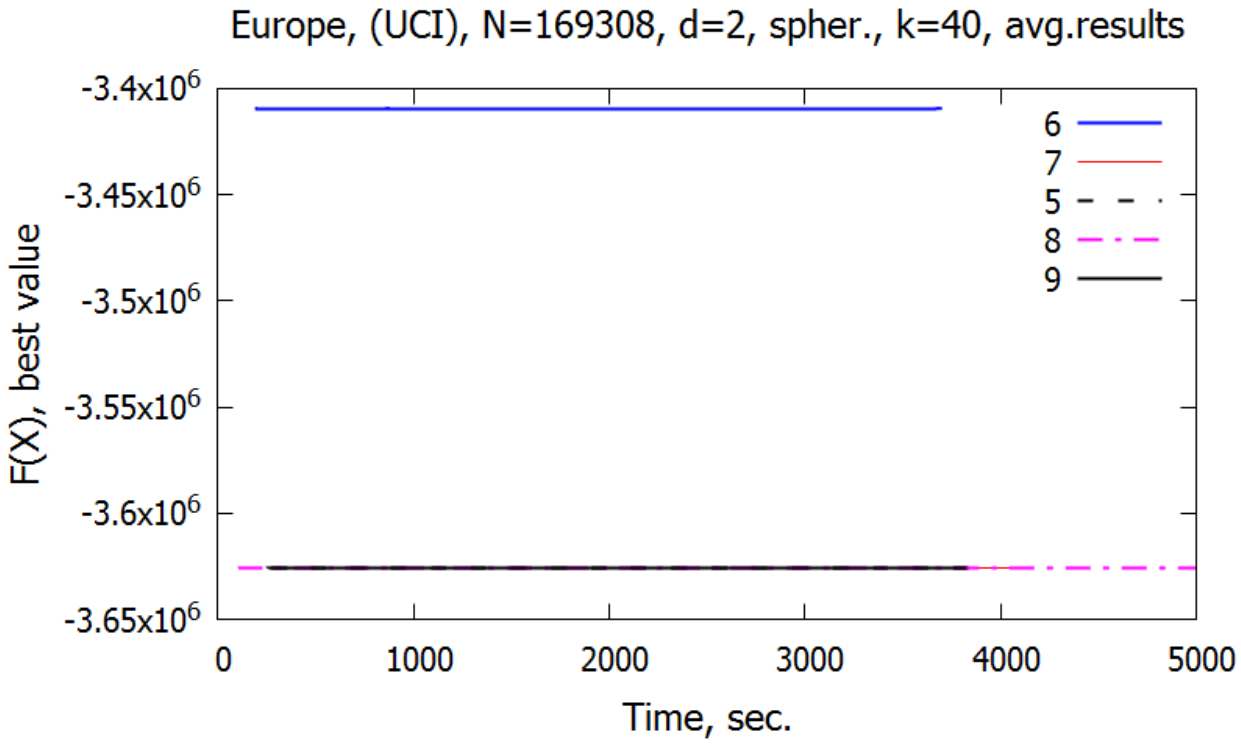
Рисунок 2.10 –Сравнение новых алгоритмов (ГА и VNS) и ЕМ-алгоритмов (классический ЕМи его модификации СЕМи SEM) для набора данных ИС1526ЛЕ2.

о оси абсцисс – время в секундах, по оси ординат – достигнутое значение целевой функции

1 – ЕМ-алгоритм (ЕМ);
2 – модификация ГА (GA-ONE);
3 – модификация ГА (GA-FULL);
4 – модификация ГА (GA-RAND);
5 – модификация VNS алгоритма(VNS-1);
6 – СЕМ-алгоритм (СЕМ);
7 – SEM-алгоритм (SEM);
5 – модификация VNS алгоритма (VNS-1);
8 – модификация VNS алгоритма (VNS-2);
9 – модификацияVNSалгоритма (VNS-3);



А)

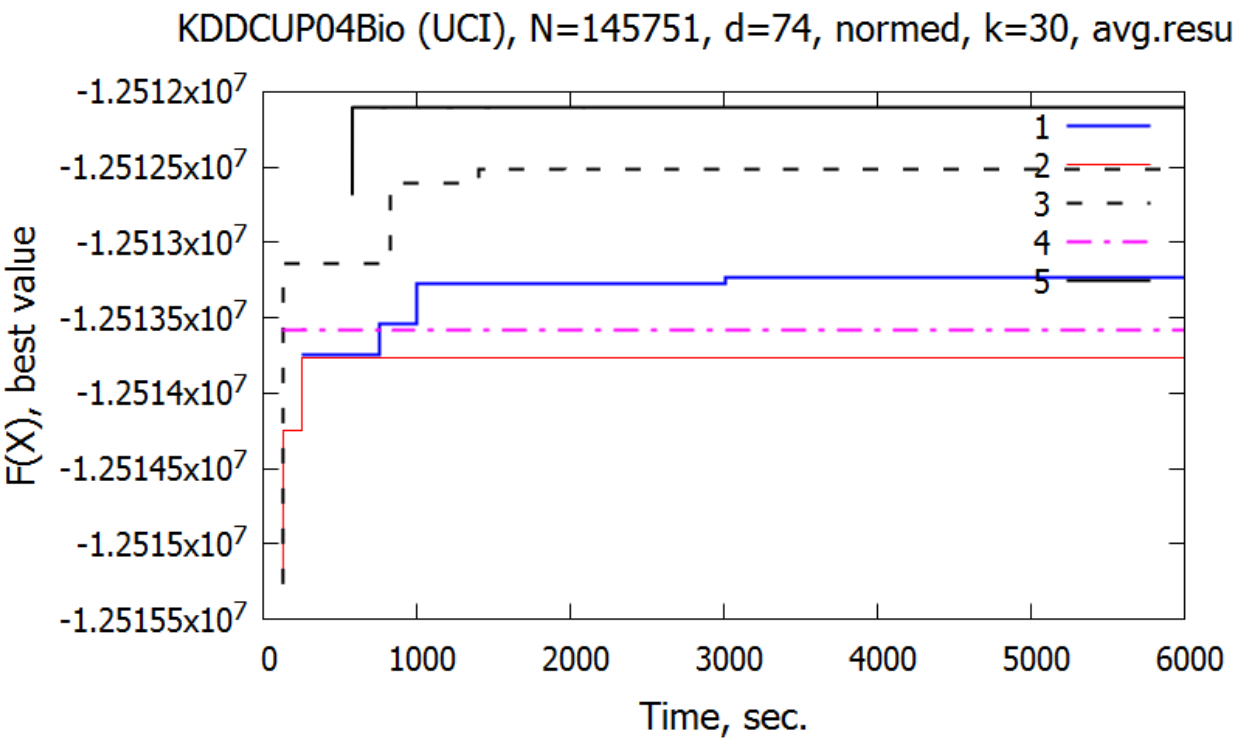


Б)

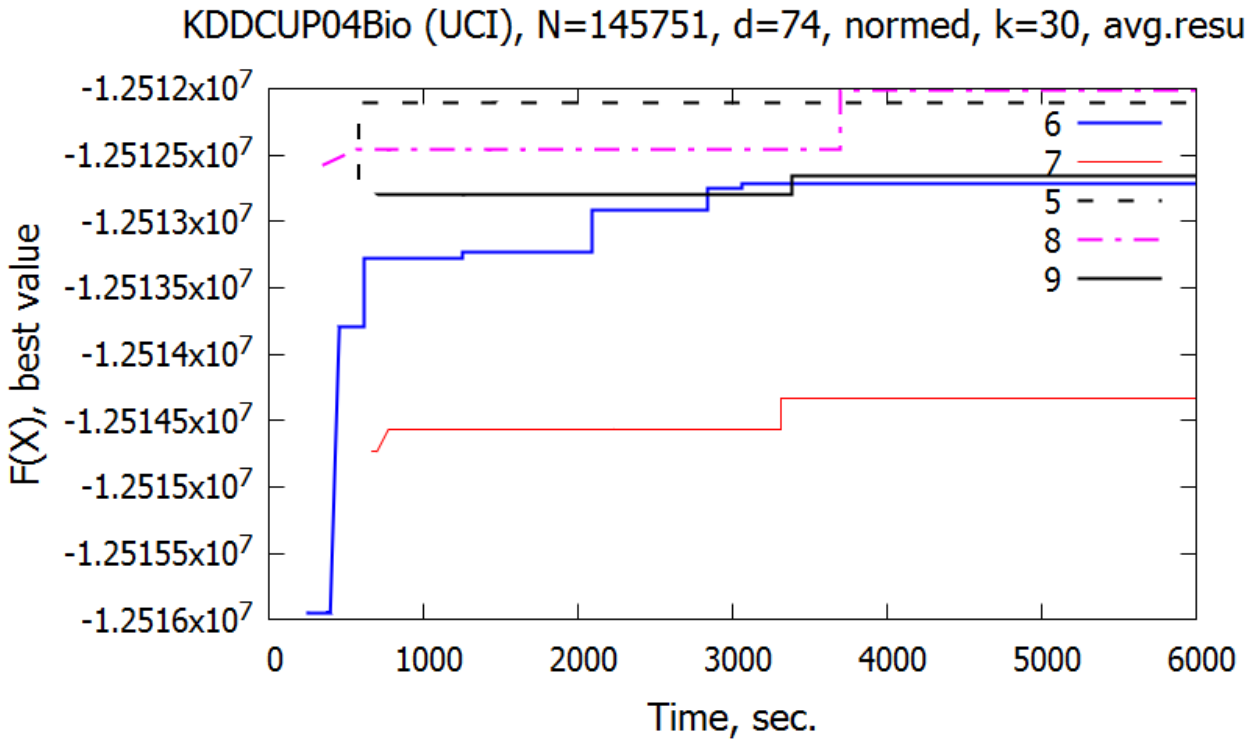
Рисунок 2.11 –Сравнение новых алгоритмов (ГА и VNS) и ЕМ-алгоритмов (классический ЕМи его модификации СЕМи SEM) для набора данных Europe (UCI).

о оси абсцисс – время в секундах, по оси ординат – достигнутое значение целевой функции.

1 – ЕМ-алгоритм (EM);
2 – модификация ГА (GA-ONE);
3 – модификация ГА (GA-FULL);
4 – модификация ГА (GA-RAND);
5 – модификация VNS алгоритма(VNS-1);
6 – СЕМ-алгоритм (CEM);
7 – SEM-алгоритм (SEM);
5 – модификация VNS алгоритма (VNS-1);
8 – модификация VNS алгоритма (VNS-2);
9 – модификацияVNSалгоритма (VNS-3);



А)

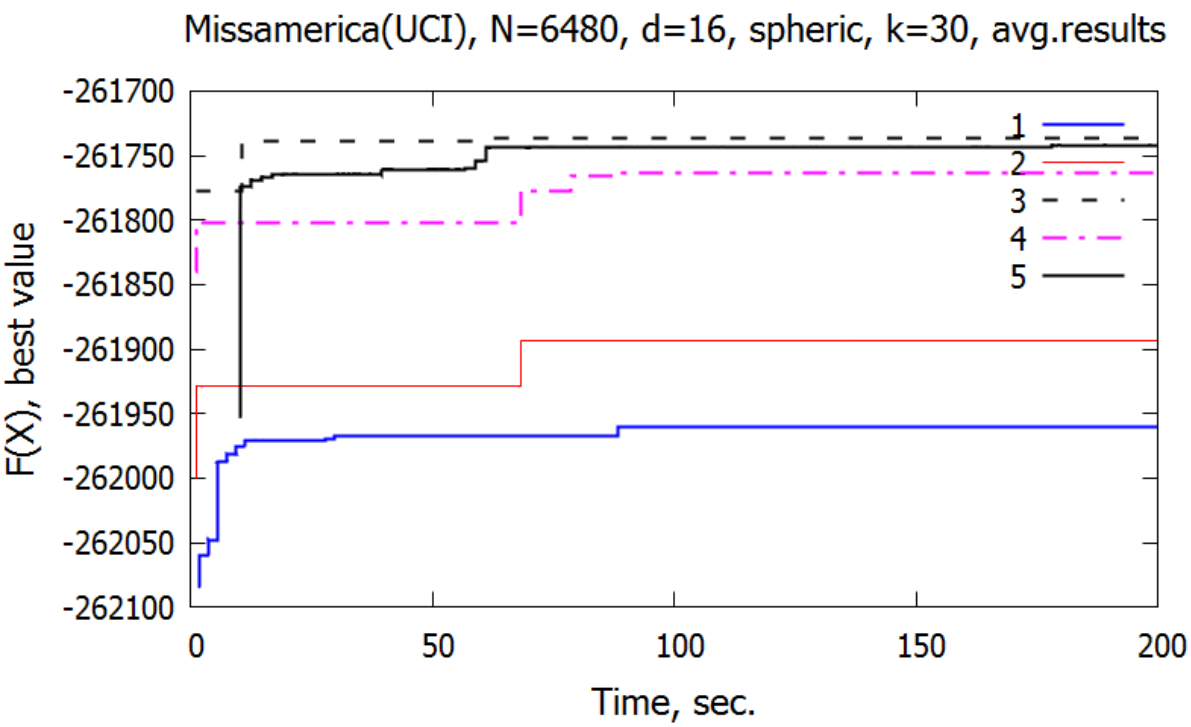


Б)

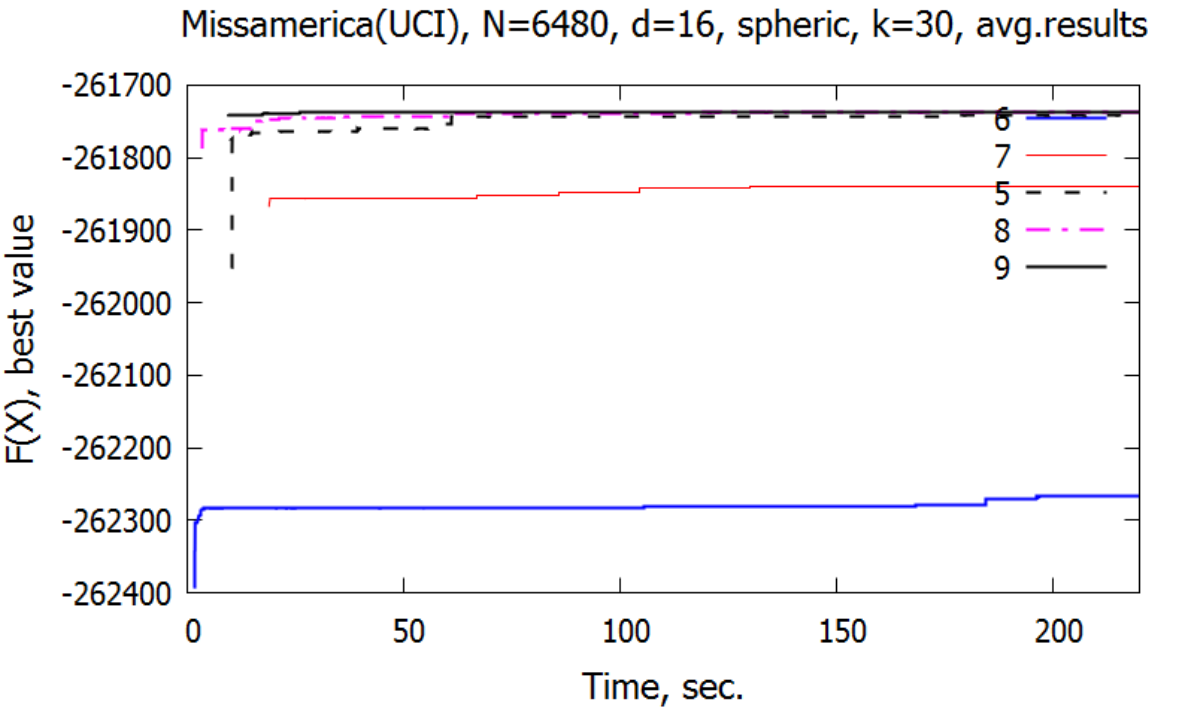
Рисунок 2.12 –Сравнение новых алгоритмов (ГА и VNS) и EM-алгоритмов (классический EMи его модификации CEMи SEM) для набора данных KDDCup04Bio (UCI).

По оси абсцисс – время в секундах, по оси ординат – достигнутое значение целевой функции.

1 – EM-алгоритм (EM);
2 – модификация ГА (GA-ONE);
3 – модификация ГА (GA-FULL);
4 – модификация ГА (GA-RAND);
5 – модификация VNS алгоритма(VNS-1);
6 – CEM-алгоритм (CEM);
7 – SEM-алгоритм (SEM);
5 – модификация VNS алгоритма (VNS-1);
8 – модификация VNS алгоритма (VNS-2);
9 – модификацияVNSалгоритма (VNS-3);



А)



Б)

Рисунок 2.13 –Сравнение новых алгоритмов (ГА и VNS) и ЕМ-алгоритмов (классический ЕМи его модификации CEMи SEM) для набора данных MissAmerica (UCI).
По оси абсцисс – время в секундах, по оси ординат – достигнутое значение целевой функции.

1 – ЕМ-алгоритм (EM);
2 – модификация ГА (GA-ONE);
3 – модификация ГА (GA-FULL);
4 – модификация ГА (GA-RAND);
5 – модификация VNS алгоритма(VNS-1);
6 – CEM-алгоритм (CEM);
7 – SEM-алгоритм (SEM);
5 – модификация VNS алгоритма (VNS-1);
8 – модификация VNS алгоритма (VNS-2);
9 – модификацияVNSалгоритма (VNS-3);

2.5 Результаты вычислительных экспериментов

Для тестирования новых алгоритмов с чередующимися окрестностями 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(www.cs.uci.edu/mlearn/mlrepository.html).

Для расчетов использовалась вычислительная система с процессором Xeon X5650 2.67 GHz, 12Gb RAM.

В таблицах 2.3 и 2.4 приведен результаты вычислительных экспериментов и выполнено сравнение новых алгоритмов с известными.

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. Новые алгоритмы неэффективны для малых задач и задач с небольшим числом кластеров.

Отметим также, что результаты новых алгоритмов имеют высокие значения среднеквадратичного отклонения, что говорит о том, что эти алгоритмы достигают хороших результатов не в каждой попытке, в связи с чем их, вероятно, также целесообразно запускать в режиме мультистарта. При этом классический EM-алгоритм даже при длительном мультистарте стабильно не позволяет даже приблизиться к значениям целевой функции, достигаемым новыми алгоритмами.

Полные результаты вычислительных экспериментов содержатся в таблицах А.1- А20 Приложения А.

Таблица 2.3 – Сравнительные результаты известных алгоритмов (модификации EM-алгоритма) 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., 0,5 мин., 2 кластера						
F(x)	3790,1∗			3665,6↓	3665,6↓	3673,7
σ	104,4			0,0	0,0	44,0
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
Mopsi (UCI), N=6014, d=2, лапласово, 10 мин., 20 кластеров						
F(x)	25628,7	29326,1	22574,5↓	31224,8	31343,6	31896,8*
σ	1334,1	882,3	14690,2	402,9	309,4	212,6
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.

Таблица 2.4 –Сравнение результатов известных алгоритмов (модификации EM-алгоритма) 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
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
Mopsi (UCI), N=6014, d=2, лапласово.,15 мин., 20 кластеров						
F(x)	28145,8	28542,1	27901,3↓	32253,4	31828,3	32304,9∗
σ	12067,6	534,2	1219,6	354,2	539,8	203,2
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.

Результаты Главы 2

Таким образом, с одной стороны, метод жадных эвристик [30, 154] может быть успешно применен для построения эффективных алгоритмов решения задач разделения смеси распределений. При этом сохраняется важное свойство алгоритмов, полученных с применением данного подхода: высокая точность получаемых результатов. Но полученные алгоритмы не обладают свойством высокой стабильности получаемых решений, характерной для алгоритмов метода

жадных эвристик, что связано с особенностями решаемых задач: известные алгоритмы локального поиска, такие как SEM-алгоритм, позволяют получать гораздо более стабильный результат, но при этом стабильно уступающий новым алгоритмам. В то же время, для некоторых практических задач, в качестве примера которых можно привести задачи автоматической группировки электрорадиоизделий [32, 151, 39], сформулированные в виде задач разделения смеси гауссовых распределений [175] результатов тестовых испытаний [80, 77], новые в ходе нескольких (не более 10) попыток запуска позволяют найти, вероятно, точный результат задачи или, по крайней мере, результат, который не получается превзойти с применением известных алгоритмов, благодаря чему данный недостаток новых алгоритмов нивелируется.

Новые алгоритмы не являются ни алгоритмами псевдобулевой оптимизации, ни алгоритмами для задач размещения, для которых разработан метод жадных эвристик, что расширяет область его использования. Кроме того, введен новый вариант стратегии глобального поиска, предусмотренной методом жадных эвристик, в виде поиска в чередующихся окрестностях одного решения путем комбинирования его элементов с элементами других промежуточных решений.

Результатом настоящей работы является расширение сферы применения метода жадных эвристик новым классом задач – задачами разделения смеси сферических и некоррелированных гауссовых распределений [69], обогащение данного метода новыми жадными эвристическими процедурами и стратегиями глобального поиска, специализированными для таких задач, благодаря чему получены новые алгоритмы, стабильно превосходящие по точности получаемых результатов известные алгоритмы для некоторых классов задач. В частности, таким классом задач являются задачи разделения смесей сферических и некоррелированных гауссовых распределений в пространствах большой размерности (десятки-сотни измерений) с числом векторов данных от сотен до десятков тысяч.

Разработанные новые эволюционные алгоритмы для задач разделения смеси гауссовых распределений позволяют достигать более точных результатов по

сравнению с классическим EM-алгоритмом, запускаемым в режиме мультистарта [70]. Новые алгоритмы дополняют область применения алгоритмов, основанных на методе жадных эвристик [30] задачами нечеткой кластеризации и подтверждают потенциальную применимость метода жадных эвристик к весьма широкому классу задач кластеризации с различными целевыми функциями.

При этом объем разделяемых данных: до сотен измерений и до сотен тысяч векторов данных.

Алгоритмы работают за конечное время (до нескольких минут), позволяющее строить на их основе интерактивные системы.

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:

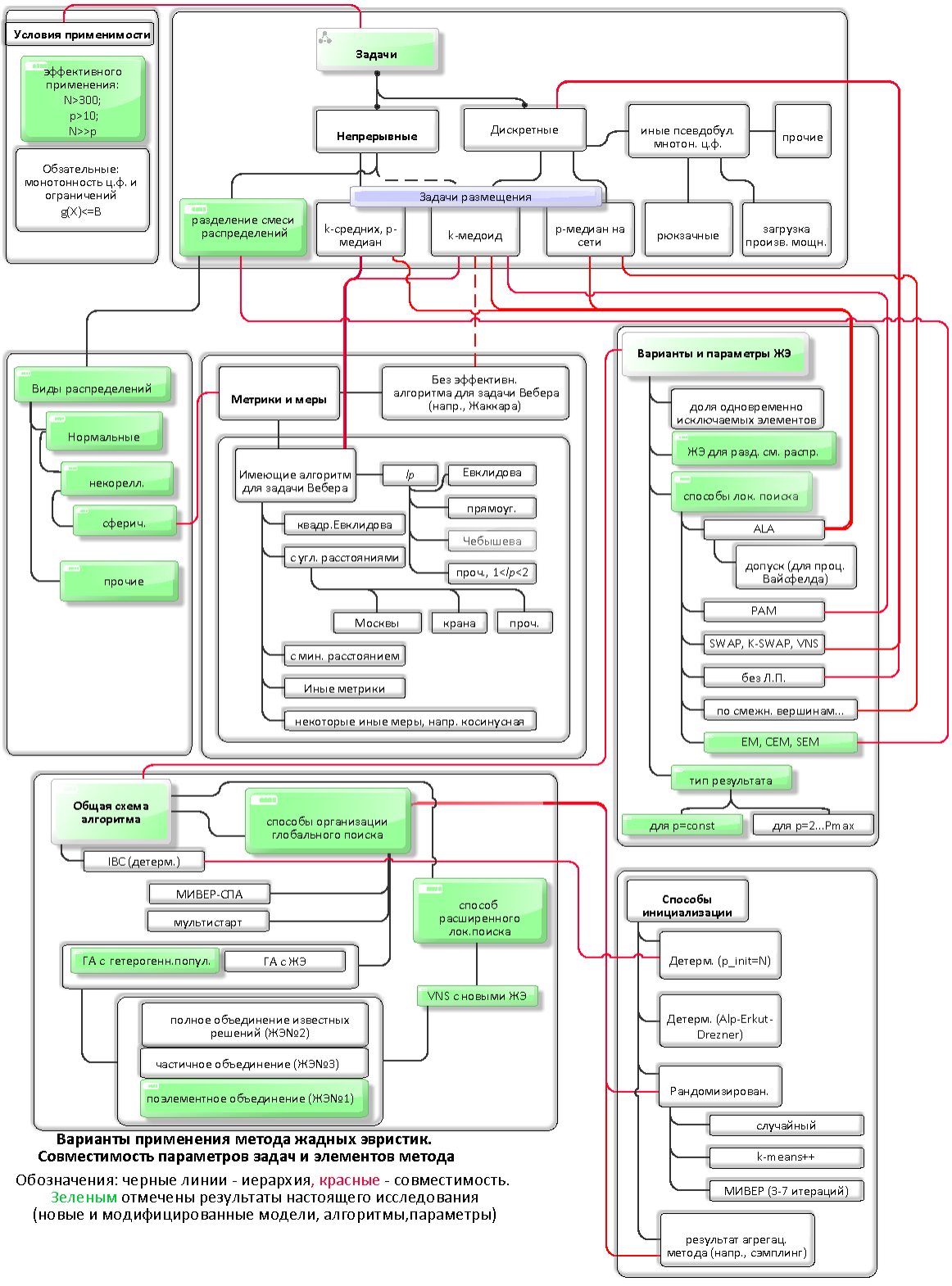
А. Мультистарт базовой жадной эвристики (Алгоритм 2.1) или VNS (Алгоритм 2.5) или ГА (Алгоритм 2.6);

Б. Одна из модификаций жадных эвристик (Алгоритм 2.2 или 2.3 или 2.4);

В. EM, CEM или SEM и т.д.

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.

Представленная на рис.2.14 схема совместимости компонентов метода жадных эвристик содержит новые компоненты (выделены зеленым), расширяющие возможности метода для задач разделения смеси распределений.



Варианты применения метода жадных эвристик.
Совместимость параметров задач и элементов метода

Обозначения: черные линии - иерархия, красные - совместимость.
Зеленым отмечены результаты настоящего исследования
(новые и модифицированные модели, алгоритмы,параметры)

Рисунок 2.14– Схема совместимости компонентов метода жадных эвристик. Зеленым выделены новые компоненты для разделения смеси распределений

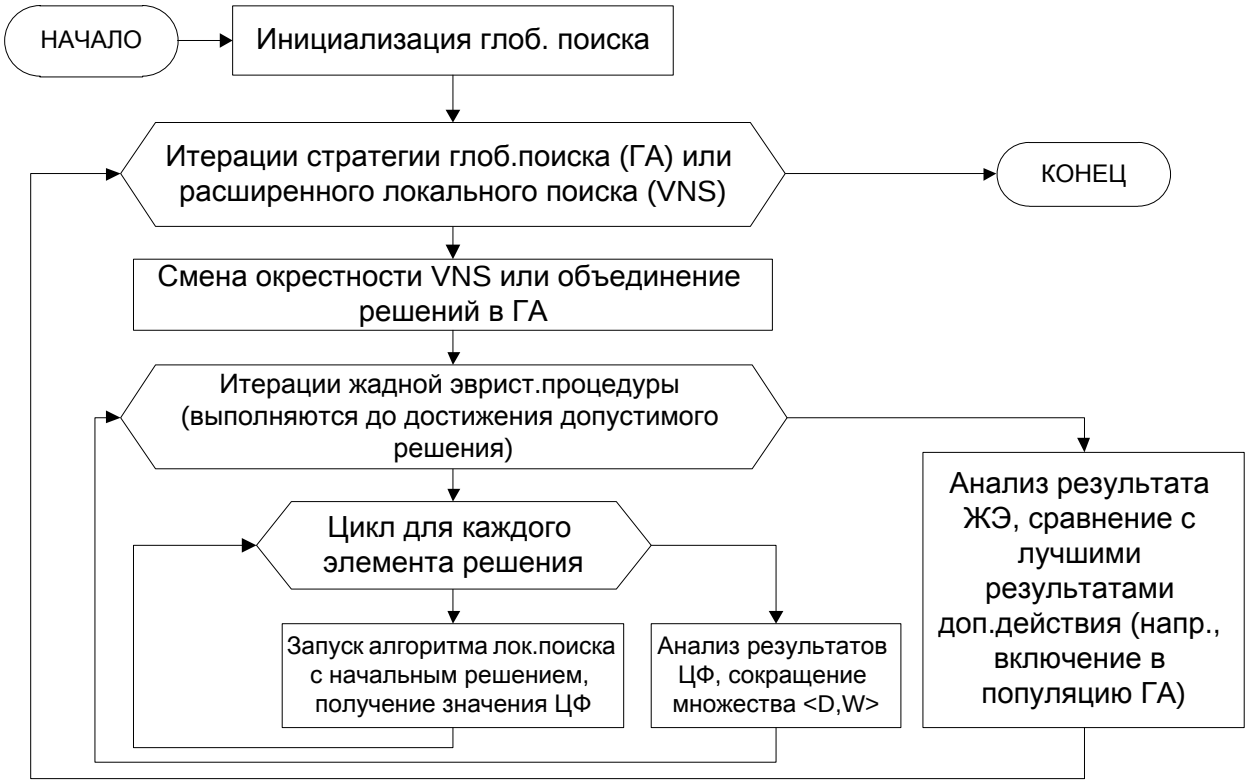


Рисунок 2.15 – Общая схема обновленного алгоритма метода жадных эвристик для разделения смеси распределений

На рис. 2.15 приведена обновленная обобщенная блок-схема организации алгоритмов метода жадных эвристик, состоящего из трех вложенных циклов:

1. *Первый цикл* предназначен для организации стратегии глобального поиска. В качестве такой стратегии могут выступать генетический алгоритм, либо VNS (в этом случае реализуется стратегия расширенного локального поиска). В этом цикле формируются промежуточные решения избыточной мощности.

2. На *втором цикле* мощность промежуточного решения приводится к допустимому значению. Для этого запускаются новые жадные процедуры.

3. *Третий цикл* для локального поиска. Мы используем ЕМ-алгоритм.

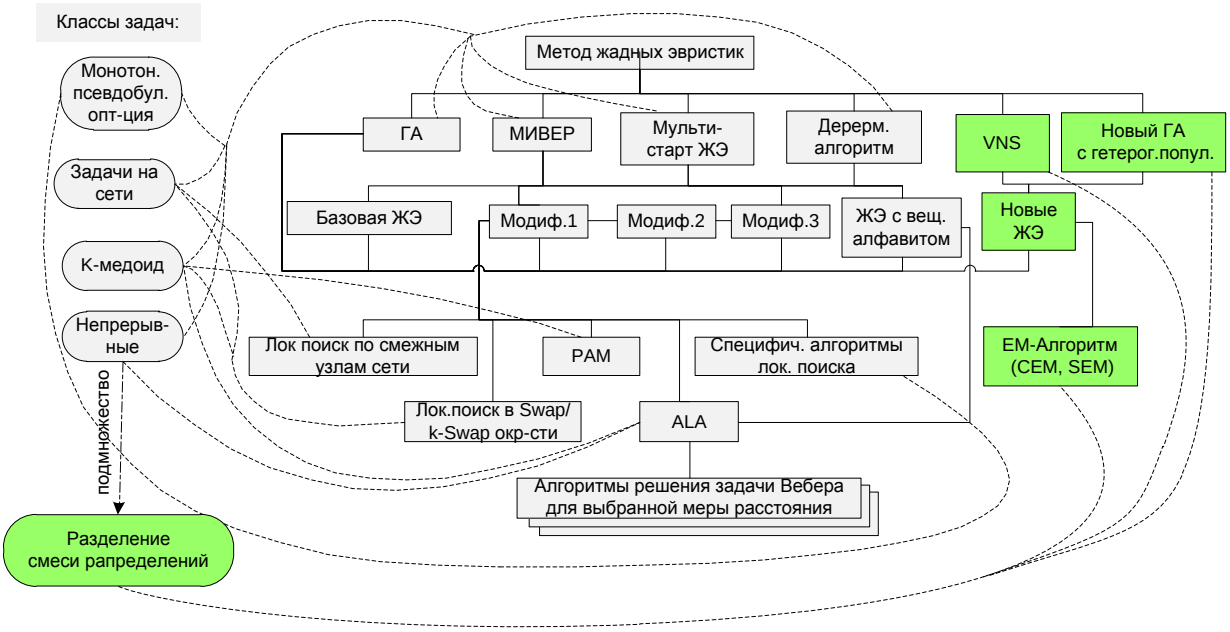


Рисунок 2.16 – Компоненты метода жадных эвристик, их взаимная совместимость (сплошные линии) и применимость к классам задач. Зеленым выделены новые компоненты для задач разделения смеси распределений

Рисунок 2.16 содержит структурную схему метода жадных эвристик, цветом выделены новые компоненты для задач разделения смеси распределений, разработанные в рамках данного диссертационного исследования. Порядок следования элементов по вертикали отражает последовательность выполнения алгоритмов.

Таким образом, в качестве основных результатов Главы 2 можно выделить следующие:

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

2. Разработаны новые алгоритмы разделения смеси распределений, а именно:

- новые генетические (эволюционные) алгоритмы

- новые алгоритмы поиска в чередующихся окрестностях.

Оба типа алгоритмов позволяют достигать более точного по значению целевой функции результата по сравнению с известными (EM-алгоритм и его

модификации).

Но VNS-алгоритмы работают несколько быстрее и более пригодны для построения интерактивных систем при большом объеме обрабатываемых данных.

Все разработанные алгоритмы являются развитием метода жадных эвристик.

ГЛАВА 3. МОДЕЛЬ И СИСТЕМА РАЗДЕЛЕНИЯ СБОРНЫХ ПАРТИЙ ИЗДЕЛИЙ ПРОМЫШЛЕННОЙ ПРОДУКЦИИ НА ОДНОРОДНЫЕ ПАРТИИ

3.1 Модель сборной партии электрорадиоизделий на основе моделей k-средних и k-медоид

В России, в отличие от США и стран Западной Европы, отсутствуют специализированные производства электронной компонентной базы (ЭКБ) - электрорадиоизделий (ЭРИ) для космической отрасли, поэтому ЭРИ общего военного (неспециализированного) применения [80, 81] категорий качества «ВП» и «ОС» («ОСМ») должны подвергаться демонстрации возможности использования в аппаратуре космических аппаратов (КА). ЭКБ иностранного производства, которая находит все более широкое распространение в аппаратуре КА, также должна подвергаться квалификации по условиям применения и уровню качества, поскольку на настоящее время не существует никаких документов о гармонизации систем качества отечественной и импортной ЭКБ.

Поэтому демонстрация возможности использования ЭРИ в аппаратуре КА в течение длительного времени (на основе разработки принципов и правил) включает разработку методологии обеспечения качества и работоспособности ЭКБ при воздействии факторов космического пространства.

Для исключения попадания в бортовую аппаратуру КА с длительными САС потенциально ненадежных ЭРИ в последние годы внедряется новый принцип комплектования аппаратуры через специализированные испытательные технические центры [43, 40] с проведением операций сплошного входного контроля ЭРИ, дополнительных отбраковочных испытаний (ДОИ), диагностического неразрушающего контроля (ДНК) с применением выборочного разрушающего физического анализа (РФА). Задачей ДОИ и ДНК ЭРИ является, по существу, индивидуальная отбраковка элементов, имеющих скрытые дефекты изготовления. РФА проводится с целью определения соответствия образцов ЭРИ

требованиям конструкции и технологического процесса изготовления и выявления нарушений этих требований.

Таким образом, все проводимые над ЭРИ испытания можно разделить на две группы:

1) сплошные испытания для всей партии элементов – ДОИ, ДНК.

2) выборочные испытания для нескольких элементов из партии – РФА.

Многие из эксплуатируемых до 2000 года КА, разработанных и изготовленных АО «Информационные спутниковые системы» имени академика М. Ф. Решетнева», в первые дни, месяцы эксплуатации имели замечания по качеству функционирования: сбои, перерывы в связи, отказы, значительная часть которых по результатам анализа возникала из-за отказов ЭРИ. И только на эксплуатируемом с 18 апреля 2000 года КА «Sesat» не выявлено существенных замечаний к ЭРИ в течение более 15 лет эксплуатации. Одной из главных причин, по мнению многих специалистов, является то, что впервые в практике все 100% ЭРИ, комплектующих бортовую аппаратуру КА «Sesat», прошли ДОИ, ДНК и РФА [40]. То, что аналогичные результаты, а именно – отсутствие или существенное снижение количества замечаний к работе ЭРИ, прошедших данный набор испытаний (ДОИ+ДНК+РФА) [55, 59], позволяет сделать вывод о состоятельности подхода к испытаниям, в состав которого входят именно эти три основных компонента.

Важное значение имеет разработка методов прогнозирования и обеспечения работоспособности ЭРИ при неблагоприятных внешних воздействиях. Одно из центральных мест при этом занимают методы обеспечения устойчивости к тепловым и радиационным нагрузкам [48, 74, 56].

Вопросы обеспечения радиационной стойкости (РС) БА изложены в обширной литературе, например, в [57, 58, 44, 49], но она, в основном, посвящена применению ЭРИ в предположении, что стойкость любого ЭРИ из производственной партии известна и одинакова. На самом деле РС ЭРИ внутри производственной партии различна и зависит от содержащихся в каждом ЭРИ внутренних дефектов (дислокации, неконтролируемых примесей, других точечных дефектов) [36].

Собственно, выявление наиболее существенных из таких дефектов в партиях изделий является целью проведения РФА. При этом распространять результаты проведенного РФА на всю поступившую партию изделий необходимо с большой осторожностью. Для этого нужно, как минимум, быть уверенными в том, что мы имеем дело действительно с единой партией ЭРИ, изготовленной из единой партии сырья. Поэтому выявление истинных производственных партий из предположительно сборных партий ЭРИ является одним из важнейших мероприятий при проведении испытаний.

Результатом диагностических испытаний, проводимых испытательным центром, является набор параметров каждого экземпляра поступившей с завода-изготовителя партии. Результат каждого из видов испытаний является числовой характеристикой (чаще всего – напряжение или ток в какой-либо цепи в тех или иных условиях). Сделать вывод об однородности либо разнородности партии изделий по условиям производства следует на основе анализа этих характеристик. Хотя к диапазону каждой из этих характеристик применяются жесткие требования [81, 75], незначительные на первый взгляд колебания сразу нескольких характеристик позволяют сделать вывод о том, что части партии произведены в разных условиях.

Для решения задачи выявления в поступившей партии экземпляров изделий, произведенных в разных условиях и, следовательно, имеющих несколько различные характеристики, был применен наиболее часто использующийся на практике метод кластерного анализа [165] – метод k-средних. Идея метода заключается в том, что на каждой итерации вычисляется центр для каждого кластера, полученного на предыдущем шаге, затем все объекты вновь разбиваются на группы в соответствие с тем, какой из новых центров оказался ближе.

В своем оригинальном исполнении метод может быть недостаточно эффективен из-за того, что результат сильно зависит от выбора исходных центров кластеров. В данной разработке были использованы поисковые методы оптимизации для оптимального выбора центров кластеров. Результаты исследований показали, что такой подход значительно повышает точность

определения групп в соответствие с выбранной метрикой, но в тоже время является приемлемым по трудоемкости в отличие от более сложных подходов.

Задачу k-средних можно отнести к задачам непрерывной теории размещения [34]. В действительности, задача сводится к нахождению k точек (центров, центроидов, главных точек) в d-мерном пространстве характеристик (здесь d – число измерений характеристик), таких, чтобы сумма расстояний от каждого из векторов данных до ближайшего к нему из k центров достигала минимума:

$$F(X_1,...,X_k) = \sum_{i-1}^{n} \min_{x \in \{x_1,...,x_k\}} \|A_i - X\|_2^2.$$

Требования качества, предъявляемые к электронным компонентам, например, в космической отрасли, настолько высоки, что определять принадлежность изделий к одной либо различным производственным партиям приходится по расхождениям (расстояниям) совокупности измерений, едва превышающим точность измерений. Таким образом, результаты каждого из измерений фактически представляют собой дискретные значения, шаг изменения которых определяется точностью измерительного прибора.

При решении практических задач автоматической группировки в пространстве характеристик в качестве меры расстояния чаще всего используется квадрат евклидовой метрики, поскольку в данном случае нахождение центра каждого из кластеров является элементарной задачей, выполняемой за один шаг, и вычислительная сложность алгоритма падает по сравнению с алгоритмом с евклидовой метрикой, при которой вычисление центра (медианы) [203] – итеративный процесс [204, 119]. Векторами данных при этом являются наборы характеристик, в нашем случае –данные результатов испытаний каждого из изделий, выраженные в виде набора числовых характеристик различной (в общем случае) размерности. Задача является задачей глобального поиска.

Задачу k-средних можно считать частным случаем p-медианной задачи. В общем случае непрерывную p-медианную задачу можно сформулировать

следующим образом:

$$\arg\min_{x_1,\ldots,X_p \in R^d} \sum_{i=1}^{N} w_i \min_{j\in\{1,p\}} L\left(X_j, A_i\right).$$

Здесь $\{A_1,...,A_N\}$ – множество известных точек – «потребителей» в d-мерном пространстве (в случае задачи k-средних именуемых векторами данных), $w_1...w_N$ – их весовые коэффициенты (чаще всего равны 1, если никаким из обрабатываемых данных не отдается предпочтение), $X_1...X_N$ – искомые точки, $L(\ )$ – некоторая функция (метрика) расстояния. Задача NP-трудная.

В качестве метода локального поиска используется ALA-процедура.

Неоднократно предпринимались попытки применения и других эвристических алгоритмов. Например, в работах [6, 89, 33, 35] используется алгоритм, основанный на методе изменяющихся вероятностей, который, однако, показал свою эффективность лишь как вспомогательный метод для ускорения сходимости генетического алгоритма на начальных итерациях. В работе Ливановой и Лореш [161] предложен алгоритм муравьиных колоний и симуляции обжига, который показал свою ограниченную эффективность.

Методы кластерного анализа – динамично развивающаяся область науки, тесно связанная с развитием искусственного интеллекта, использует методы теории размещения и аналогичные методы. Если классические исследования теории размещения направлены на получение точных результатов, то исследования, связанные с развитием методов кластерного анализа, как правило, сосредоточены на повышении скорости работы соответствующих алгоритмов. В то же время, в некоторых случаях требуются быстрые и при этом весьма точные (не обязательно в строгом смысле) методы. К таким случаям относится и наша задача, применяемый способ решения которой, как и всякой задачи в космической отрасли, требует строгой регламентации.

3.2 Нормирующий коэффициент в алгоритмах классификации и автоматической группировки

Для выявления однородных партий используются алгоритмы кластеризации, некоторые подходы к этой проблеме рассмотрены в [151, 32]. При этом для корректной работы алгоритма необходима нормирование данных, полученных в результате испытаний – приведение их к единой шкале.

В данной работе предложен новый подход, в основе которого лежит особый способ нормирования значений характеристик ЭРИ, полученных в результате ДОИ и ДНК, с применением граничных значений дрейфа (изменения) этих характеристик, выявляемых в ходе ДНК.

Наиболее распространенные способы нормирования – 0-1-нормирование по минимальному и максимальному значению, когда минимальное значение показателя в выборке принимается равным 0, максимальное – равным 1, соответственно, остальные показатели располагаются в интервале (0;1) [83]. Другой способ – по среднеквадратичному отклонению:

$$a_{i,k} = \frac{a^*_{i,k} - \overline{a^*_k}}{\sigma(a^*_k)}.$$

Здесь $a^*_{i,k}$ - значение k-го показателя i-го вектора ненормированных данных, $\overline{a^*_k}$ – среднее значение k-го показателя по выборке, $\sigma(a^*_k)$ – его среднеквадратичное отклонение.

Указанные выше способы нормирования не сильно отличаются друг от друга, и оба они имеют существенный изъян, который проявляет себя при обработке реальных данных. Изделия разных партий одного и того же типа могут иметь различный разброс некоторого параметра. Так, для одной партии этот параметр может изменяться в широком диапазоне, и значение этого параметра на самом деле влияет на группировку изделий. А для изделий другой партии тот же параметр может быть стабилен, и фактически он не должен тогда влиять на группировку, но в соответствие с приведенными выше способами нормирования

малейшее его изменение может повлиять на результат группировки. Для того чтобы этого избежать, требуется выявить, насколько изменения каждого параметра рассматриваемой партии являются значимыми.

Мы использовали способ нормирования, позволяющий получать гораздо более точное разбиение на производственные партии по данным тестовых испытаний электрорадиоизделий. Способ основан на использовании установленных границ дрейфа параметров изделий.

Дрейф – изменение параметров ЭРИ в ходе эксплуатации, а также изменение этих параметров в ходе проведения ДНК, имитирующего жесткие условия эксплуатации. Основным элементом ДНК [151] является электротермотренировка – имитация эксплуатации ЭРИ под нагрузкой в условиях высоких температур в течение нескольких часов или суток. После электротермотренировки проводится повторный замер параметров и сравнение с их исходными значениями. Разность значений после ДНК и исходных значений мы и будем называть дрейфом в данной работе.

Дрейф параметров на заводах-изготовителях ЭРИ не оценивается. При проведении электротермотренировки фиксируются случаи выхода параметров за нормы ТУ. Фактически это означает, что отбраковываются ЭРИ с дрейфом параметров, превышающим границы X_{TY}. В качестве параметра, характеризующего запас параметрической надежности по величине дрейфа параметров, предлагается использовать коэффициент, определяемый соотношением:

$$k_{Д} = 1/\left[(X_{TY} - M(x_{TY})/(X_{D} - M(y)\right],$$

где $M(X_{TY})$ - оценка математического ожидания параметра в соответствии с ТУ;

M(y) - оценка математического ожидания дрейфа параметра в конкретной производственной партии;

Y_{D} - граница дрейфа параметра, установленная по выборке;

X_{TY} - верхняя или нижняя граница параметра по ТУ.

Математическое ожидание параметра по ТУ, фактически, является номинальным значением параметра ЭРИ по ТУ и может быть приближенно оценено по формуле:

$$M(X_{TY}) = \frac{1}{2}(X_{TY1} - X_{TY2}),$$

где X_{TY1}и X_{TY2}– соответственно верхняя и нижняя граница параметра по ТУ.

В качестве параметра, характеризующего нормы на границы дрейфа, установленные по выборке, принимается величина:

$$Y_D = M(y) \pm k \cdot s_D,$$

где s_D – оценка среднеквадратического отклонения дрейфа параметра;

k - толерантный коэффициент.

$M(y)$ определяется как $M(y) = \frac{1}{m}\sum_{i=1}^{M} Y_i$.

Соответственно, $s_D = \sqrt{1/m\sum_{i=1}^{m}[Y_i - M(y)]^{2`}}$.

Оценка допустимых границ дрейфа производится по данным испытаний достаточно крупной партии ЭРИ.

Ниже (таблица 3.1) приведены результаты оценки коэффициентов по вышеприведенной методике для партии ИС 597СА3 в количестве 988 штук.

Установленные таким образом нормы параметров в дальнейшем используются при проведении отбраковочных испытаний. Но также они являются важными ориентирами при оценке значимости разброса параметров конкретной партии изделий и могут быть использованы для более точного нормирования данных.

Таблица 3.1 - Оценка границ дрейфа параметров по данным отбраковочных испытаний

Параметр	Значение по ТУ	Ужесточ. нормы	Нормы дрейфа	Коэфф. дрейфа
	Верхнее значение	Верхнее значение	Верхнее значение	
	Нижнее значение	Нижнее значение	Нижнее значение	
Ток потребления от отриц. источника не более, мА	1,0000	0,5856	0,0428	0,09
	-	0,5242	-0,0429	
Напряжение смещ. нуля Iio, В	0,0050	0,0021	0,0027	0,55
	-0,0050	0,0019	-0,0028	
Выход.напр-е. низкого уровня не более, В	0,4000	0,3099	0,0098	0,05
	-	0,2945	-0,0095	
Выход.напр-е высок. уровня не менее, В	-	9,4111	0,2304	0,06
	7,2000	9,0416	-0,2467	

Итак, в основе нашего способа нормирования исходных данных лежит использование допустимых границ дрейфа характеристик (показателей). В ходе электротермотренировки – одного из способов испытаний электрорадиоизделий – измеряемые показатели практически неизбежно меняются под воздействием высоких температур и значительных нагрузок. Для каждого из показателей установлены предельное значение этого изменения, называемого дрейфом. Обозначим нижнюю и верхнюю границу дрейфа k-го показателя соответственно δ_k^{min} и δ_k^{max}. Тогда предварительная обработка данных будет осуществляться согласно следующему выражению:

$$a_{i,k} = \frac{a_{i,k}^{*} - \overline{a_k^{*}}}{\delta_k^{max} - \delta_k^{min}}.$$

Как показали вычислительные эксперименты, такой способ нормирования по границам дрейфа дает разбиение по производственным партиям с гораздо меньшим количеством ошибок.

С целью проверки работоспособности метода автоматической группировки электрорадиоизделий по производственным партиям мы исследовали партии интегральных схем, заведомо составленные из двух или трех различных производственных партий. Применялись способы нормирования показателей по среднеквадратичному отклонению и по границам дрейфа. Для визуального представления многомерных данных использовалось многомерное шкалирование [24]. На диаграмме многомерного шкалирования цифрами обозначены номера предполагаемых производственных партий, «X» и «?» - не классифицированные элементы. На графике критерия силуэта показана зависимость значения критерия от предполагаемого числа производственных партий.

На рис. 3.1 показано разбиение сборной партии операционных усилителей на две предполагаемые производственные партии с использованием нормирования данных по среднеквадратичному отклонению, а также значения критерия силуэта [183] для различных вариантов разбиения на разное число партий. Видно, что критерий силуэта в данном случае не информативен, его максимум не соответствует реальному числу производственных партий (равному двум).

На рис. 3.2 даны аналогичные результаты с применением нормирования по границам дрейфа. Видно, что максимальное значение критерия силуэта четко указывает на число производственных партий, равное двум. Также уменьшилось до нуля число неклассифицированных элементов, не отнесенных ни к одной из партий. Отметим, что в обоих случаях разбиение всех элементов, кроме помеченных как ненадежно классифицированные, совпадает с их реальной принадлежностью к той или иной производственной партии. К ненадежно классифицированным отнесены элементы, лежащие за пределами кластеров.

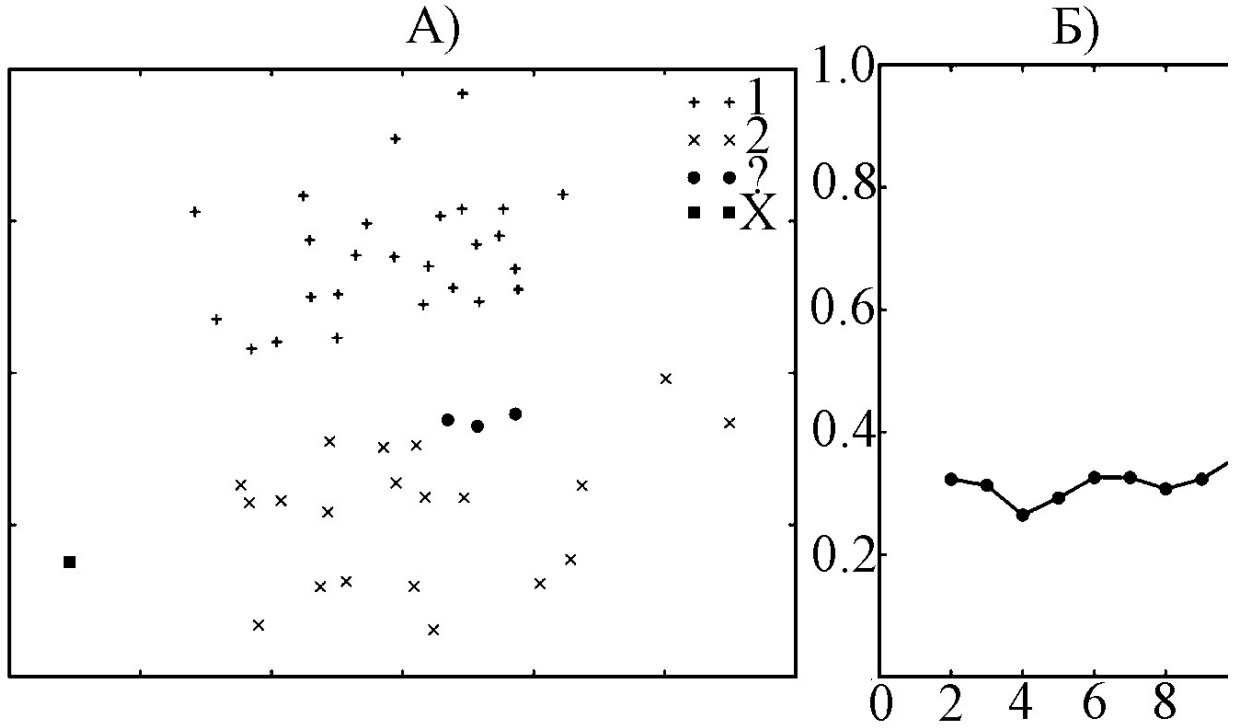


Рисунок 3.1 - Сборная партия интегральных схем 140УД17АВК, нормирование по среднеквадратичному отклонению (способ 1) А)– многомерное шкалирование; Б)– критерий силуэта (горизонтальная ось – количество кластеров, вертикальная ось – критерий силуэта)

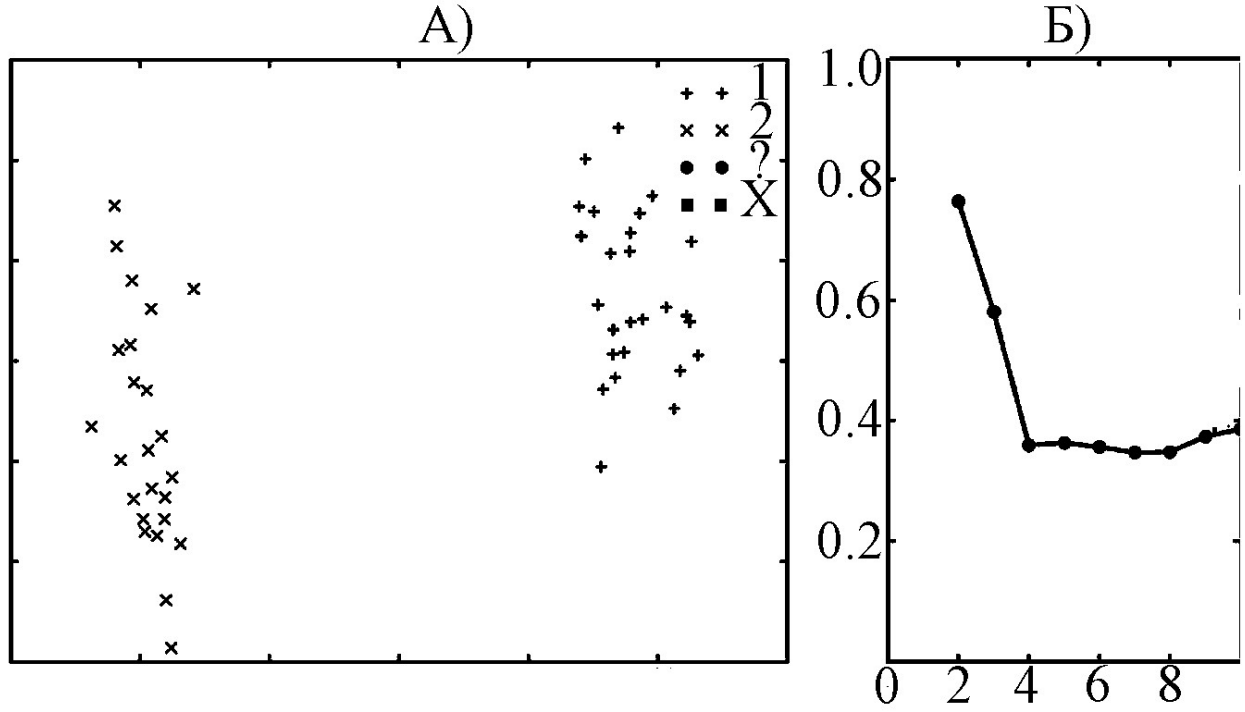


Рисунок 3.2 - Сборная партия ИС 140УД17АВК – нормирование по границам дрейфа (способ 2)А)– многомерное шкалирование; Б)– критерий силуэта (горизонтальная ось – количество кластеров, вертикальная ось – критерий силуэта)

В качестве параметров, по которым производилась классификация ЭРИ на рис. 3.2 были использованы значения параметров, снятые после электротермотренировки. Сами по себе значения дрейфа, вычисленные по данным тестирования, проведенного до и после электротермотренировки, могут рассматриваться как дополнительные параметры ЭРИ. В этом случае число параметров, по которым производится классификация ЭРИ, удваивается: к параметрам, измеренным после электротермотренировки, добавляется такое же количество параметров, значениями которых являются величины дрейфа соответствующих параметров.

Как показано на рис. 3.3, такой подход (добавление значений дрейфа в качестве дополнительных параметров) не дает преимуществ перед использованием только тех значений параметров, которые сняты после электротермотренировки. На рис. 3.3 показаны результаты классификации сборной партии ЭРИ, состоящей фактически из двух производственных партий, с использованием трех различных способов:

Способ 1. По значениям параметров, снятых после электротермотренировки, с нормированием по среднеквадратичному отклонению;

Способ 2. По значениям параметров, снятых после электротермотренировки, с нормированием по границам дрейфа;

Способ 3. По значениям параметров, снятых после электротермотренировки, и значениям дрейфа, с нормированием по границам дрейфа.

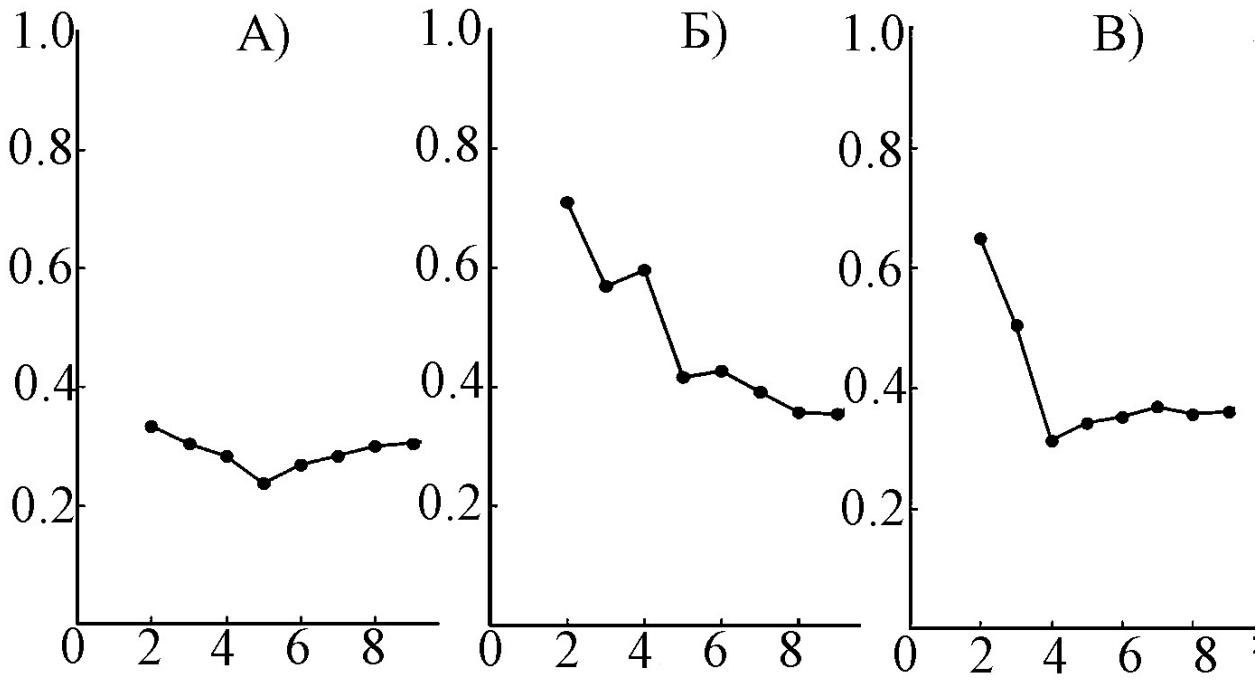


Рисунок 3.3 - Разбиение сборной партии ИС 140УД25АС1ВК с различными способами нормирования (горизонтальная ось – количество кластеров, вертикальная ось – критерий силуэта) А) –первый способ, Б) –второй способ, В) – третий способ.

Как видно из рис. 3.3, способ 3 не дает преимуществ перед способом 2 – максимум значения критерия силуэта в точке, соответствующей истинному количеству производственных партий (2 партии), менее выражен. Классификация с использованием только значений дрейфа в качестве параметров дает еще менее качественный результат.

При этом наиболее важным вопросом остается адекватность классификации. В экзаменационных выборках ИС, результаты разбиения которых показаны на рис. 3.2 и 3.3, состав элементов с разбивкой на производственные партии был заранее известен. В таблице 3.2 показаны результаты разбиения этих выборок. Видно, что способ 2 дает меньший процент неклассифицированных элементов.

Таблица 3.2 - Результаты разбиения экзаменационных выборок ЭРИ на производственные партии различными способами

Показатель	ИС 140УД25 АС1ВК, способ 1	ИС 140УД25 АС1ВК, способ 2	ИС 140УД17АВК, способ 1	ИС 140УД17 АВК, способ 2
Партия №1	30		26	
Партия №2	26		24	
Всего	56		50	
По результатам работы алгоритма классификации:				
Партии №1	22	27	17	26
Партия №2	22	26	24	24
Неклассифицированные	12 (21,4%)	3 (8,9%)	9 (18%)	0 (0%)
Ошибочные	0	0	0	0

Таким образом, предложенная система нормирования, позволяющая производить выявление однородных производственных партий в сборной партии электрорадиоизделий космического применения, обладает следующими свойствами:

а) работа системы не требует дополнительных испытаний: данные дополнительных отбраковочных испытаний и дополнительного неразрушающего контроля достаточны для выявления однородных производственных партий в сборной партии;

б) благодаря применению критерия силуэта в совокупности с особым способом нормирования данных, основанном на границах дрейфа, система позволяет определять число производственных партий в сборной партии.

Использование способа нормирования, основанного на установленных границах дрейфа параметров, позволяет производить более точную классификацию электрорадиоизделий по производственным партиям по сравнению с традиционными способами нормирования.

3.3 Моделирование сборной партии электрорадиоизделий в виде смеси многомерных гауссовских распределений по результатам неразрушающих испытаний

Предполагается, что результаты разрушающего физического анализа (РФА) распространяются на всю исследуемую партию ЭРИ. Такой подход имеет основания лишь в том случае, если партия ЭРИ изготовлена как единая производственная партия и из единой партии сырья. В противном случае различия в составе и технологии изготовления сырья (поры, трещины, вкрапления), а также возможные нарушения технологии изготовления отдельных производственных партий ставят под сомнение результаты РФА, проведенного над сборной партией ЭРИ, если состав этой партии не исследовался.

Пусть сборная партия из N ЭРИ состоит из k однородных партий объемом $N_1,N_2,\ldots N_k$ штук соответственно, при этом одна из партий (без потери общности предположим, что первая) – дефектная, которая не должна пройти РФА. Предположим, что для РФА случайным образом выбираются $N_{РФА}$ штук изделий.

Тогда вероятность того, что для РФА не будет отобран хотя бы один экземпляр первой дефектной партии, и РФА сборной партии ошибочно покажет отсутствие дефектов, составит

$$P_{\text{РФА1}} = \prod_{i=1}^{N_{\text{РФА}}} \left(1 - \frac{N_1}{N-i+1}\right).$$

Как правило, из партии для РФА отбираются 3 изделия. Тогда при объеме сборной партии 40 единиц, состоящей из трех партий по 10, 15 и 15 штук, вероятность того, что ни один элемент из первой партии не будет подвержен РФА, составит

$$P_{\text{РФА1}} = \left(1 - \frac{10}{40}\right)\left(1 - \frac{10}{39}\right)\left(1 - \frac{10}{38}\right) = 41{,}1\%.$$

Таким образом, разбор партии на однородные составляющие позволяет существенно увеличить эффективность проводимого РФА. В приведенном выше примере можно было бы обойтись без увеличения общего количества изделий,

подвергаемых РФА, отобрав по одному изделию из каждой партии, и при этом снизить вероятность непрохождения РФА дефектной партией до нуля.

Таким образом, если существует метод разделения сборных партий ЭРИ с вероятностью ошибки $P_{ош,}$ то вероятность непрохождения РФА дефектной однородной партией при $N_{РФА}=k$ составит

$$P_{\text{РФА1}} = \frac{N_i}{N} P_{\text{ош}} \quad ,$$

что для приведенного выше примера для уровня ошибок классификации 10% составит 2,5%.

Таким образом, разработка методов разделения сборной партии представляет высокую ценность при проведении РФА, даже если эти методы работают с высоким уровнем ошибок.

Расчеты, проводимые по результатам неразрушающих тестов ЭРИ, позволяют осуществлять разделение сборных партий с достаточной точностью.

Результаты тестов, проводимых 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.

Имея результаты испытаний заведомо однородной партии, изготовленной из однородной партии сырья, можно оценить характер вероятностного распределения [61] каждого из измерений (параметров) электрорадиоизделий в партии.

Некоторые результаты по оценке характера распределений достигнуты в [77]. Так, из рисунков 3.1 и 3.2 видно, что распределения отраженных на соответствующих графиках величин по своей форме близки к гауссовым (нормальным) распределениям.

Аналогичным образом, в [77] показано, что значения дрейфа параметров, полученные исходя из измерений до и после электротермотренировки, имеют вид, близкий к нормальному (см. рис.3.3).

Ниже представлены результаты исследования характера распределений

параметров однородных партий электрорадиоизделий на примере ИС 1526ТЛ1. Были исследованы изделия из трех различных однородных партий объемом 199, 535, 98 штук. Исследованию подвергнуты все полученные в ходе неразрушающих испытаний измерения. Значения каждого измерения были сгруппированы в 7 интервалов.

Результаты исследования одной из партий в таблицу 3.3.

Таблица 3.3 - Результаты исследования вероятностных распределений отдельных измерений (параметров) ИС 1526ТЛ1 на примере партии из 199 штук

№ параметра	Среднее значение	Средне-квадр. откло-нение	Критерий эксцесса	Интервалы значений параметра			
				Границы интервала		Частота попадания в интервал	Теоретическая частота попадания в интервал в предположении о нормальности распределения
T2	-0,03464	0,000355	-0,669	-∞	-0,0353	2	12,1112
				-0,0353	-0,03509	27	22,24904
				-0,03509	-0,03487	34	38,67464
				-0,03487	-0,03466	45	47,19217
				-0,03466	-0,03444	34	40,42747
				-0,03444	-0,03423	23	24,31193
				-0,03423	-0,03401	22	10,26153
				-0,03401	+∞	12	3,772006
T3	-0,03464	0,000351	-0,6329	-∞	-0,0353	1	11,80688
				-0,0353	-0,03509	25	22,21356
				-0,03509	-0,03487	36	38,97081
				-0,03487	-0,03466	48	47,66004
				-0,03466	-0,03444	31	40,63498
				-0,03444	-0,03423	26	24,15167
				-0,03423	-0,03401	19	10,00465
				-0,03401	+∞	13	3,557402
T5	-0,03465	0,000351	-0,62321	-∞	-0,0353	1	11,88415
				-0,0353	-0,03509	27	22,35092
				-0,03509	-0,03487	32	39,14586
				-0,03487	-0,03466	49	47,74116
				-0,03466	-0,03444	32	40,54652
				-0,03444	-0,03423	27	23,97934
				-0,03423	-0,03401	18	9,87299
				-0,03401	+∞	13	3,479068
T12	-1E-08	1E-07	96,95	-∞	-1E-06	2	4,08E-18
				-1E-06	-8,6E-07	0	8,66E-13
				-8,6E-07	-7,1E-07	0	2,47E-08
				-7,1E-07	-5,7E-07	0	9,55E-05

Продолжение таблицы 3.3

				-5,7E-07	-4,3E-07	0	0,051495
				-4,3E-07	-2,9E-07	0	4,038977
				-2,9E-07	-1,4E-07	0	49,57459
				-1,4E-07	0	197	145,3348
T15		1,41E-07	45,94	-∞	-1E-06	4	1,06E-08
				-1E-06	-8,6E-07	0	5,25E-06
				-8,6E-07	-7,1E-07	0	0,000949
				-7,1E-07	-5,7E-07	0	0,063484
				-5,7E-07	-4,3E-07	0	1,587031
				-4,3E-07	-2,9E-07	0	15,01864
				-2,9E-07	-1,4E-07	0	54,49758
	-2E-08			-1,4E-07	0	195	127,8323
T21		2,93E-07	17,52	-∞	-2E-06	1	9,77E-08
				-2E-06	-1,6E-06	0	0,000331
				-1,6E-06	-1,1E-06	0	0,14289
				-1,1E-06	-7,1E-07	5	8,288913
				-7,1E-07	-2,9E-07	0	70,58328
				-2,9E-07	1,43E-07	185	97,15819
				1,43E-07	5,71E-07	0	22,0597
	5,03E-09			5,71E-07	0	8	0,766687
T33		0,000263	10,96	-∞	-0,001	3	0,063677
				-0,001	-0,00071	0	1,922663
				-0,00071	-0,00043	0	19,40391
				-0,00043	-0,00014	0	66,01423
				-0,00014	0,000143	185	76,74505
				0,000143	0,000429	0	30,54554
				0,000429	0,000714	0	4,117627
	4,02E-05			0,000714	+∞	11	0,187302
T67			0,0109	-∞	2,15	1	2,434425
				2,15	2,255714	7	10,78887
				2,255714	2,361429	14	31,52724
				2,361429	2,467143	43	54,04478
				2,467143	2,572857	63	54,38786
				2,572857	2,678571	44	32,13177
				2,678571	2,784286	21	11,13618
	2,521256	0,141528		2,89	+∞	6	2,548892
T69		0,140769	0,140	2,17	2,17	1	3,110214
				-∞	2,274286	8	12,57832
				2,274286	2,378571	15	34,21205
				2,378571	2,482857	55	55,05061
				2,482857	2,587143	55	52,43781
				2,587143	2,691429	38	29,56636
				2,691429	2,795714	20	9,860443
	2,525327			2,795714	+∞	7	2,184191
T76		0,306458	-0,6699	-∞	-8,17	1	6,834989
				-8,17	-7,96286	15	18,2802
				-7,96286	-7,75571	31	38,50211
				-7,75571	-7,54857	40	52,22657
				-7,54857	-7,34143	49	45,6357
				-7,34143	-7,13429	38	25,68493
				-7,13429	-6,92714	21	9,30728
	-7,50854			-6,92714	+∞	4	2,528227

Продолжение таблицы 3.3

T89			-0,409	-∞	0,454	1	9,507756
				0,454	0,457429	13	21,31082
				0,457429	0,460857	34	40,35249
				0,460857	0,464286	64	50,74762
				0,464286	0,467714	33	42,39339
				0,467714	0,471143	27	23,52172
				0,471143	0,474571	18	8,665089
	0,464492	0,005267		0,474571	+∞	9	2,501107
T110		0,005258	-0,6934	-∞	-0,493	2	5,924274
				-0,493	-0,49	27	12,87856
				-0,49	-0,487	11	26,71273
				-0,487	-0,484	22	40,35693
				-0,484	-0,481	36	44,4137
				-0,481	-0,478	55	35,60628
				-0,478	-0,475	35	20,79314
	-0,48159			-0,475	+∞	11	12,31439
T143		0,000829	0,1044	-∞	0,0122	1	5,575501
				0,0122	0,012829	12	19,21635
				0,012829	0,013457	34	44,21939
				0,013457	0,014086	46	58,78908
				0,014086	0,014714	64	45,1765
				0,014714	0,015343	31	20,05787
				0,015343	0,015971	5	5,139224
	0,014098			0,015971	+∞	6	0,826082
T154		0,011065	-0,692	-∞	4,73	12	7,381403
				4,73	4,735714	0	12,95928
				4,735714	4,741429	39	24,6027
				4,741429	4,747143	0	35,98084
				4,747143	4,752857	60	40,53894
				4,752857	4,758571	0	35,18796
				4,758571	4,764286	61	23,53026
	4,752613			4,764286	+∞	27	18,81862

Графически результаты статистической обработки данных, приведенный в таблице 3.3 вместе с результатами обработки данных второй партии из 535 изделий того же типа представлены на рисунках 3.4-3.17.

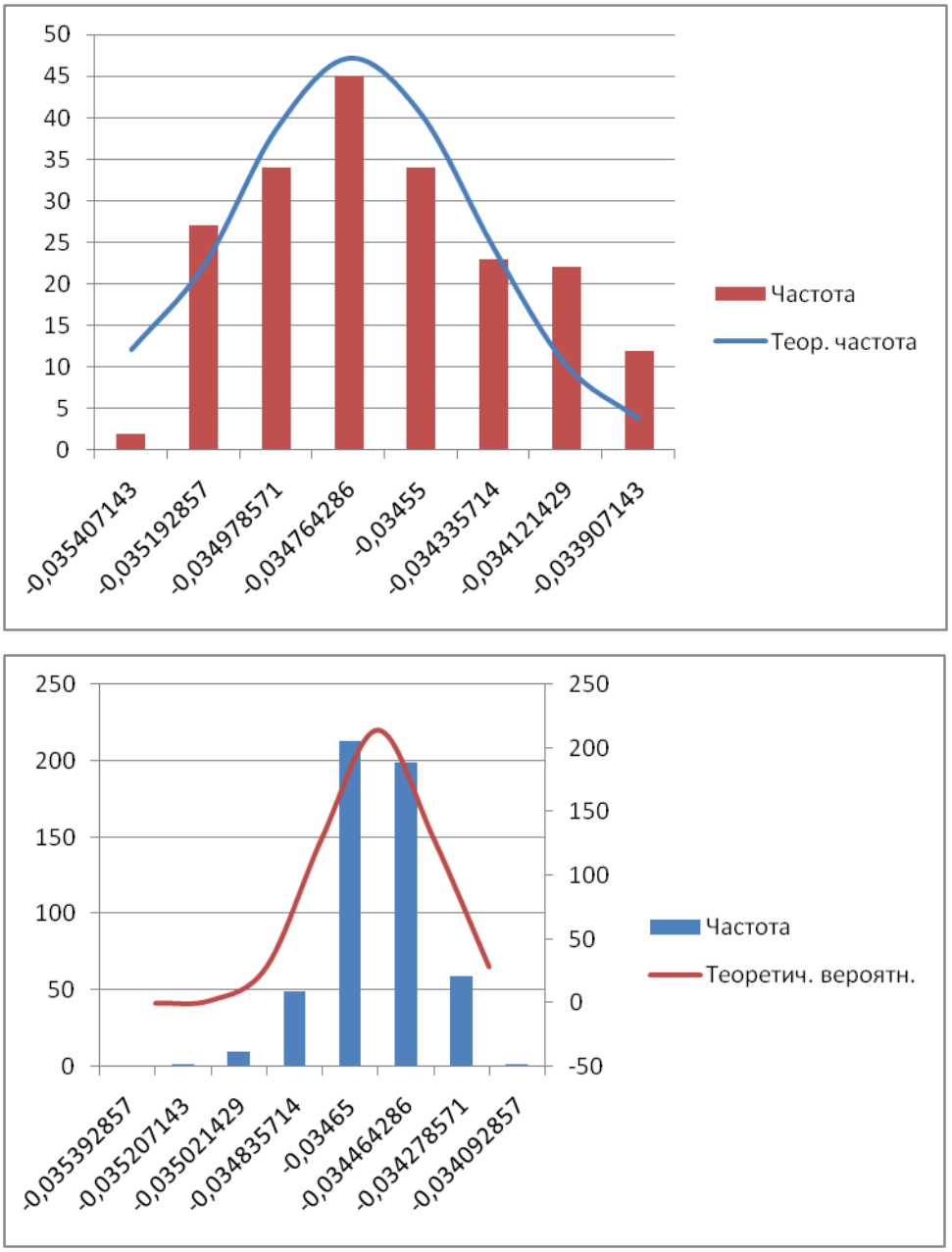


Рисунок 3.4 –Частотная гистограмма параметра Т2 ИС 1526ТЛ1. Вверху – партия №1 (199шт.), внизу – партия №2 (535 шт.)

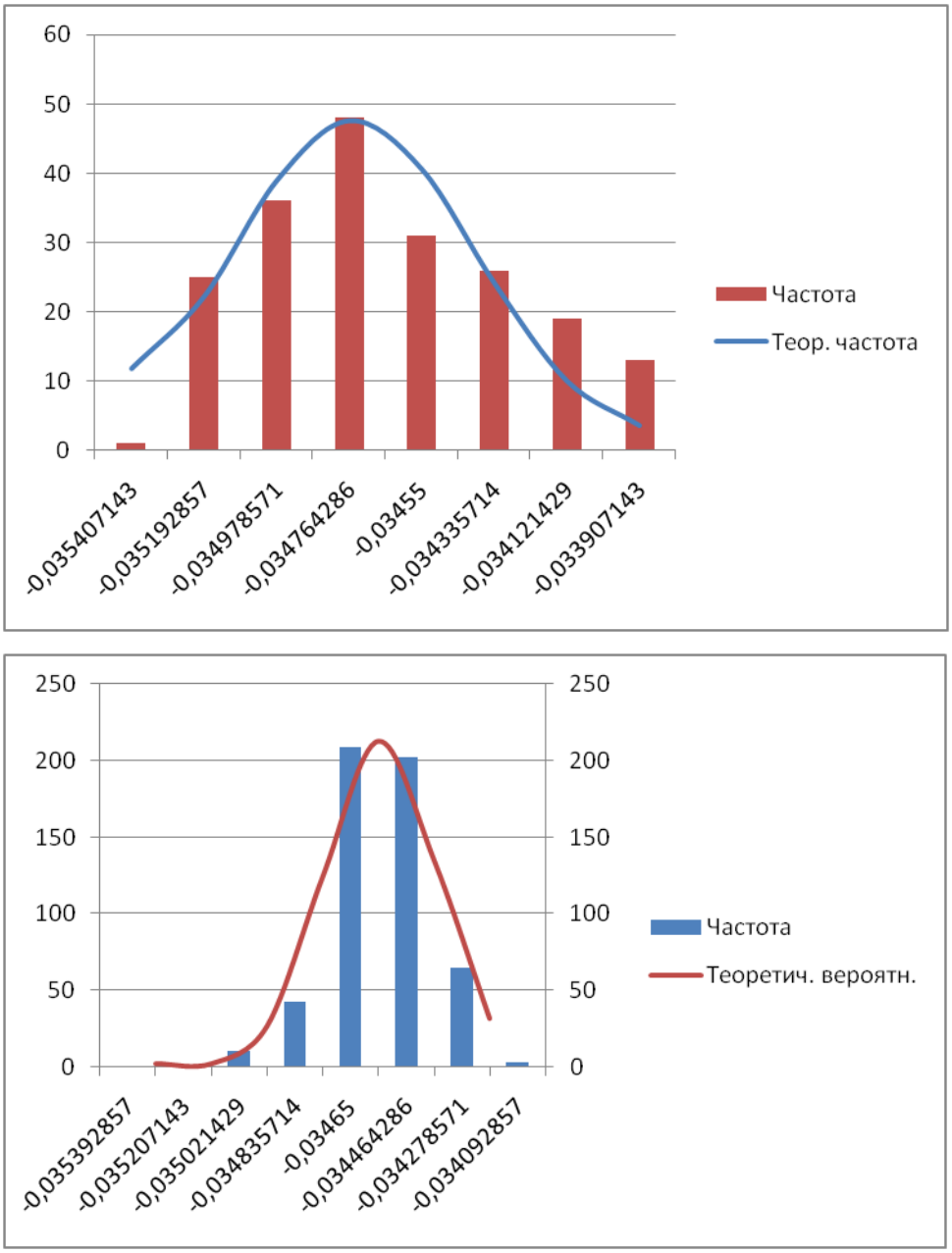


Рисунок 3.5 –Частотная гистограмма параметра Т3 ИС 1526ТЛ1. Вверху – партия №1 (199шт.), внизу – партия №2 (535 шт.)

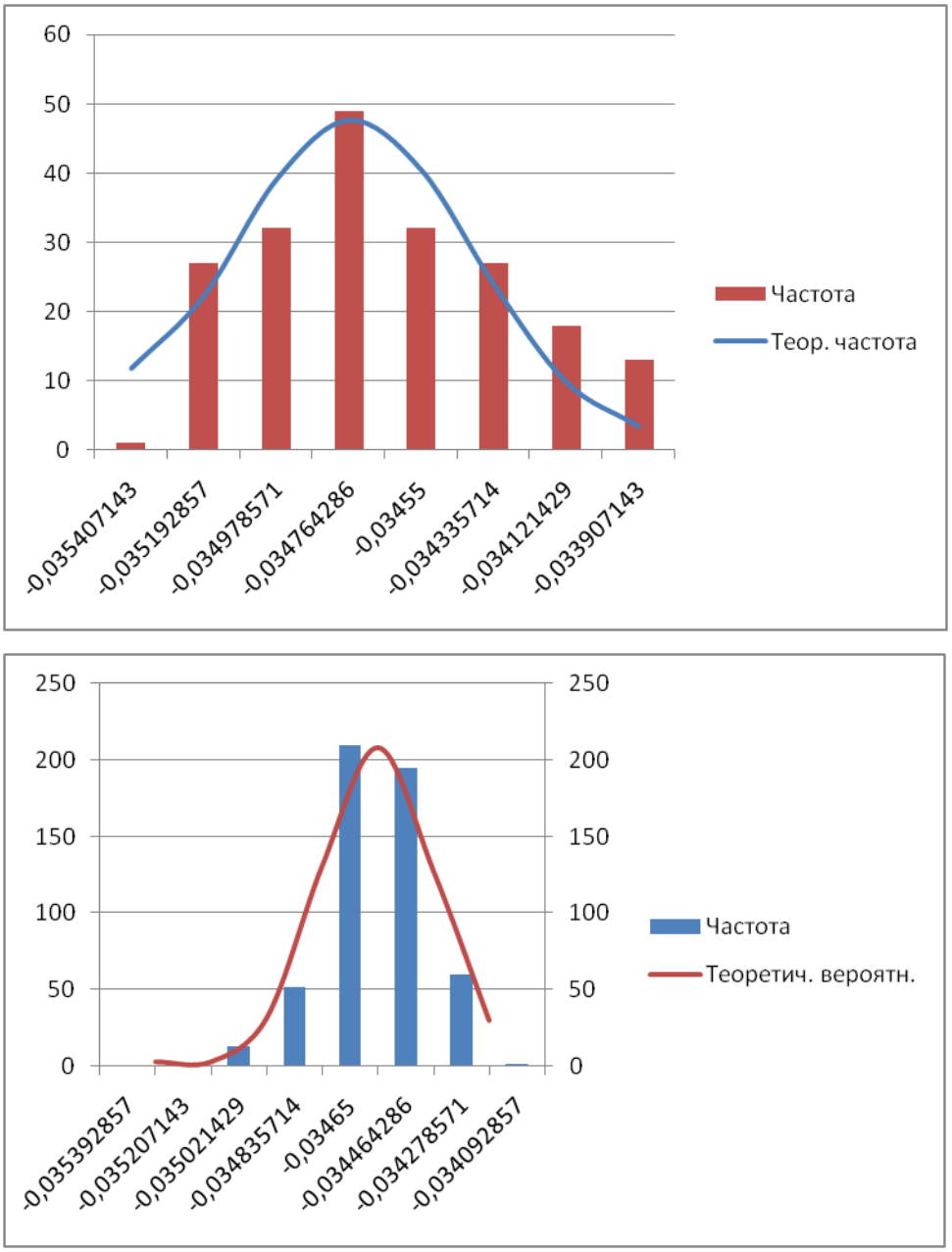


Рисунок 3.6 –Частотная гистограмма параметра Т5 ИС 1526ТЛ1. Вверху – партия №1 (199шт.), внизу – партия №2 (535 шт.)

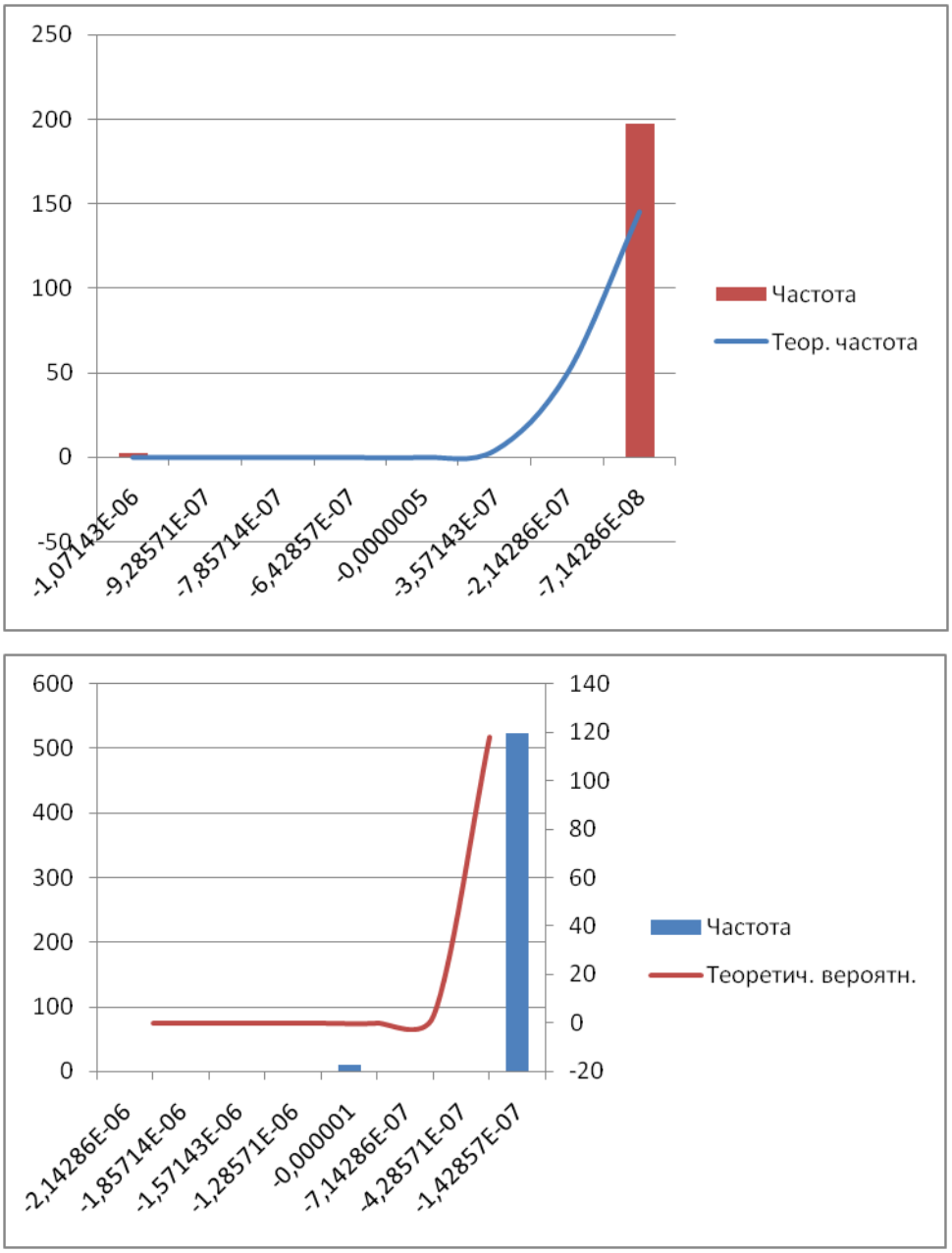


Рисунок 3.7 –Частотная гистограмма параметра Т12 ИС 1526ТЛ1. Вверху – партия №1 (199шт.), внизу – партия №2 (535 шт.)

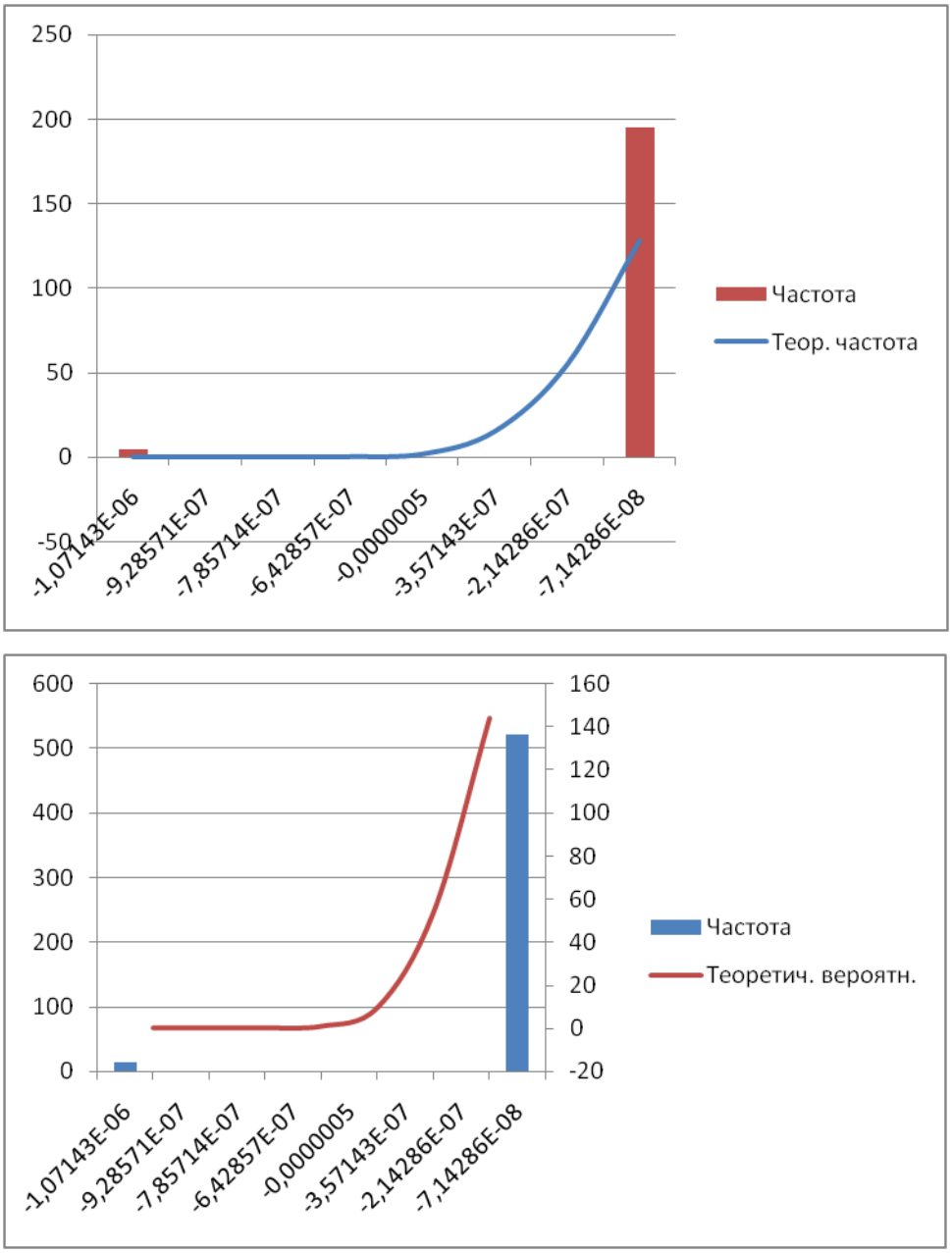


Рисунок 3.8 –Частотная гистограмма параметра Т12 ИС 1526ТЛ1. Вверху – партия №1 (199шт.), внизу – партия №2 (535 шт.)

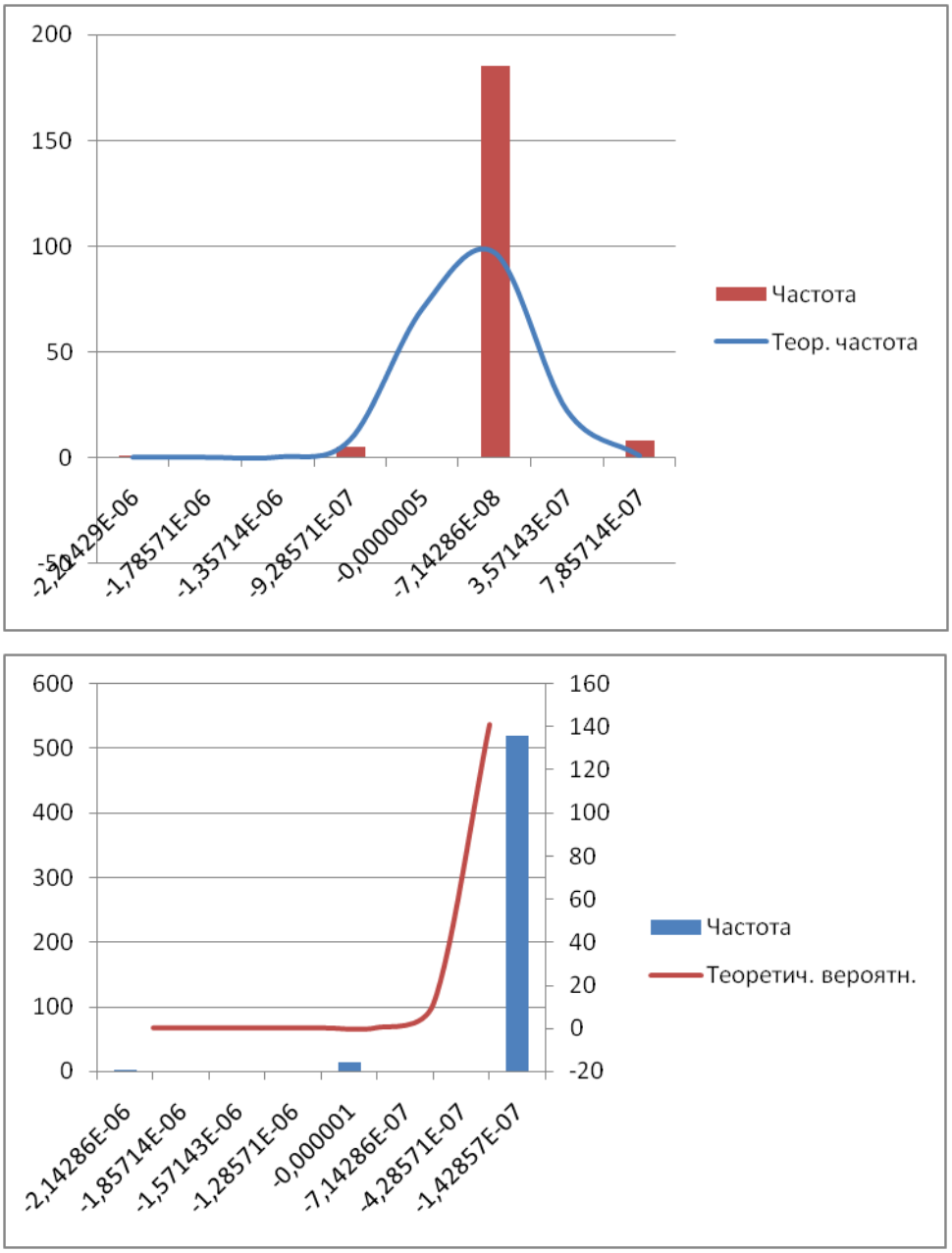


Рисунок 3.9 –Частотная гистограмма параметра T21 ИС 1526ТЛ1. Вверху – партия №1 (199шт.), внизу – партия №2 (535 шт.)

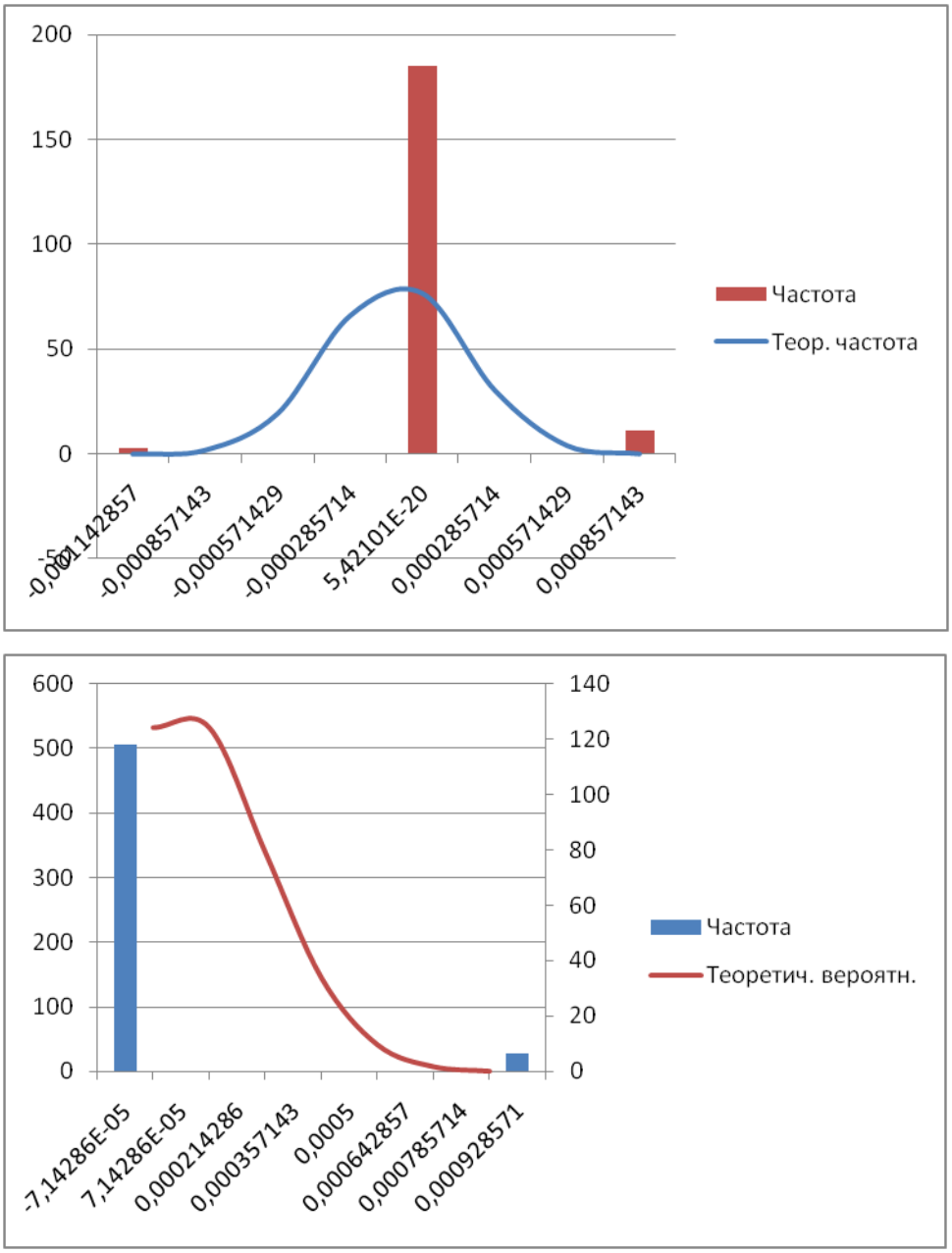


Рисунок 3.10 –Частотная гистограмма параметра Т33 ИС 1526ТЛ1. Вверху – партия №1 (199шт.), внизу – партия №2 (535 шт.)

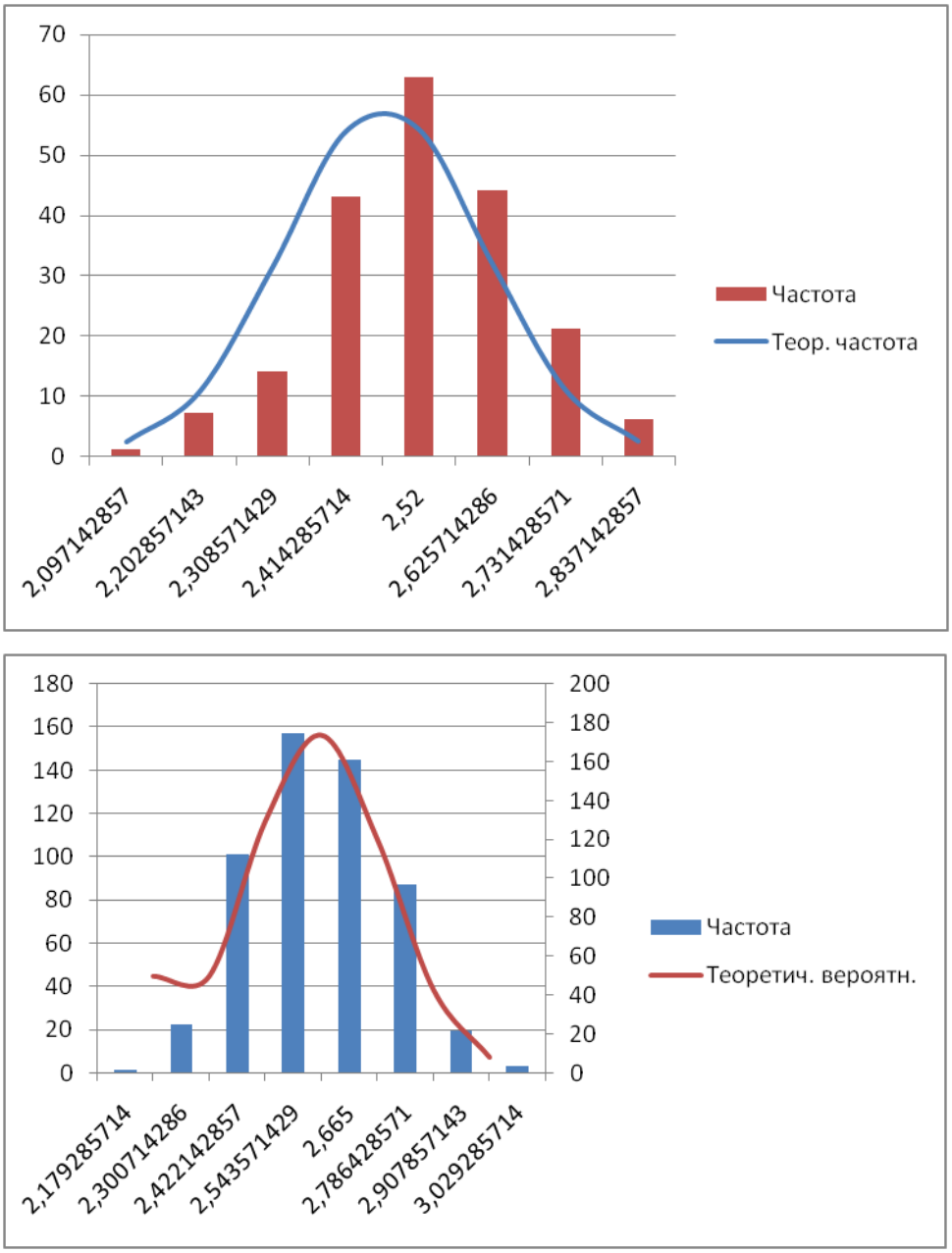


Рисунок 3.11–Частотная гистограмма параметра Т67 ИС 1526ТЛ1. Вверху – партия №1 (199шт.), внизу – партия №2 (535 шт.)

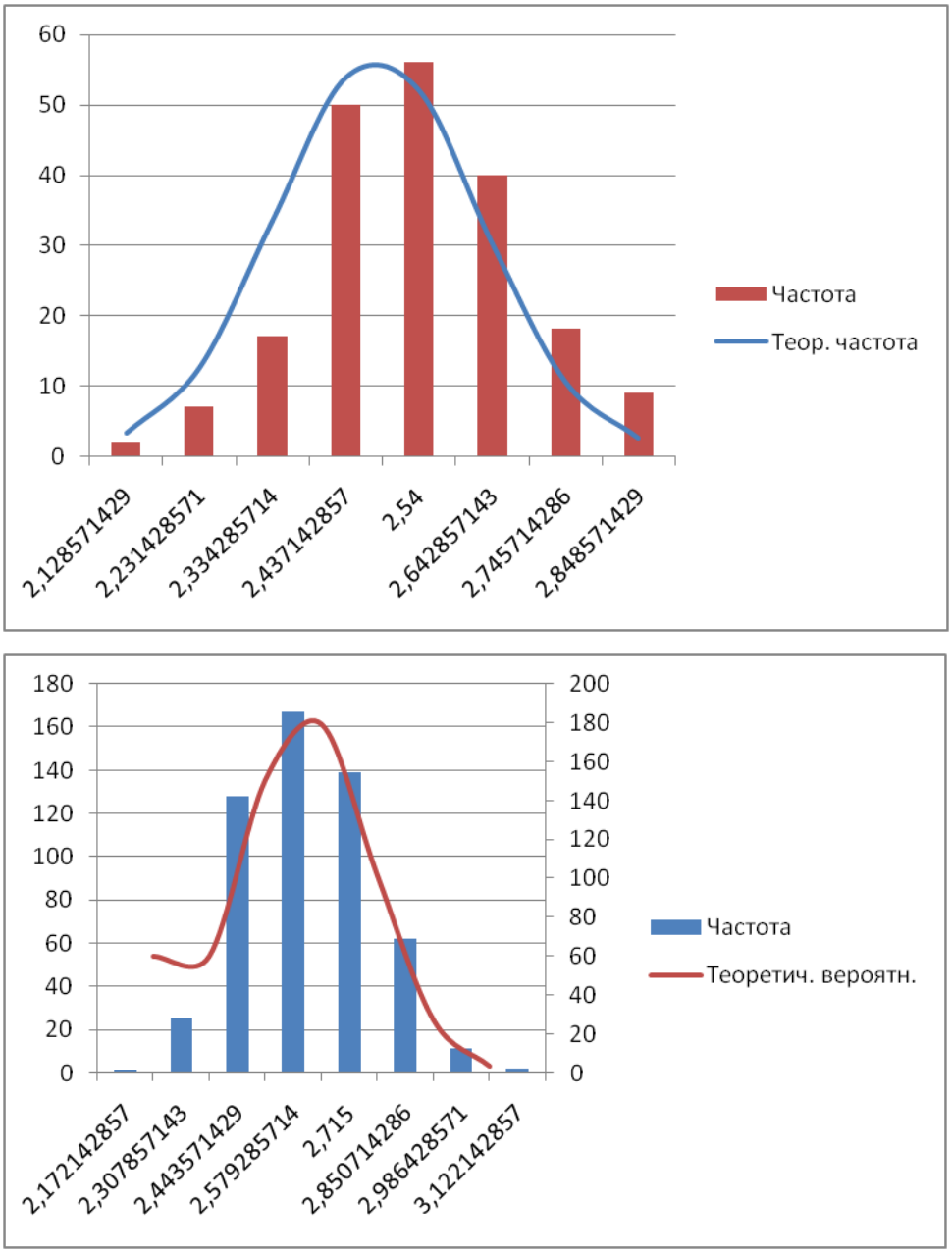


Рисунок 3.12 –Частотная гистограмма параметра Т69 ИС 1526ТЛ1. Вверху – партия №1 (199шт.), внизу – партия №2 (535 шт.)

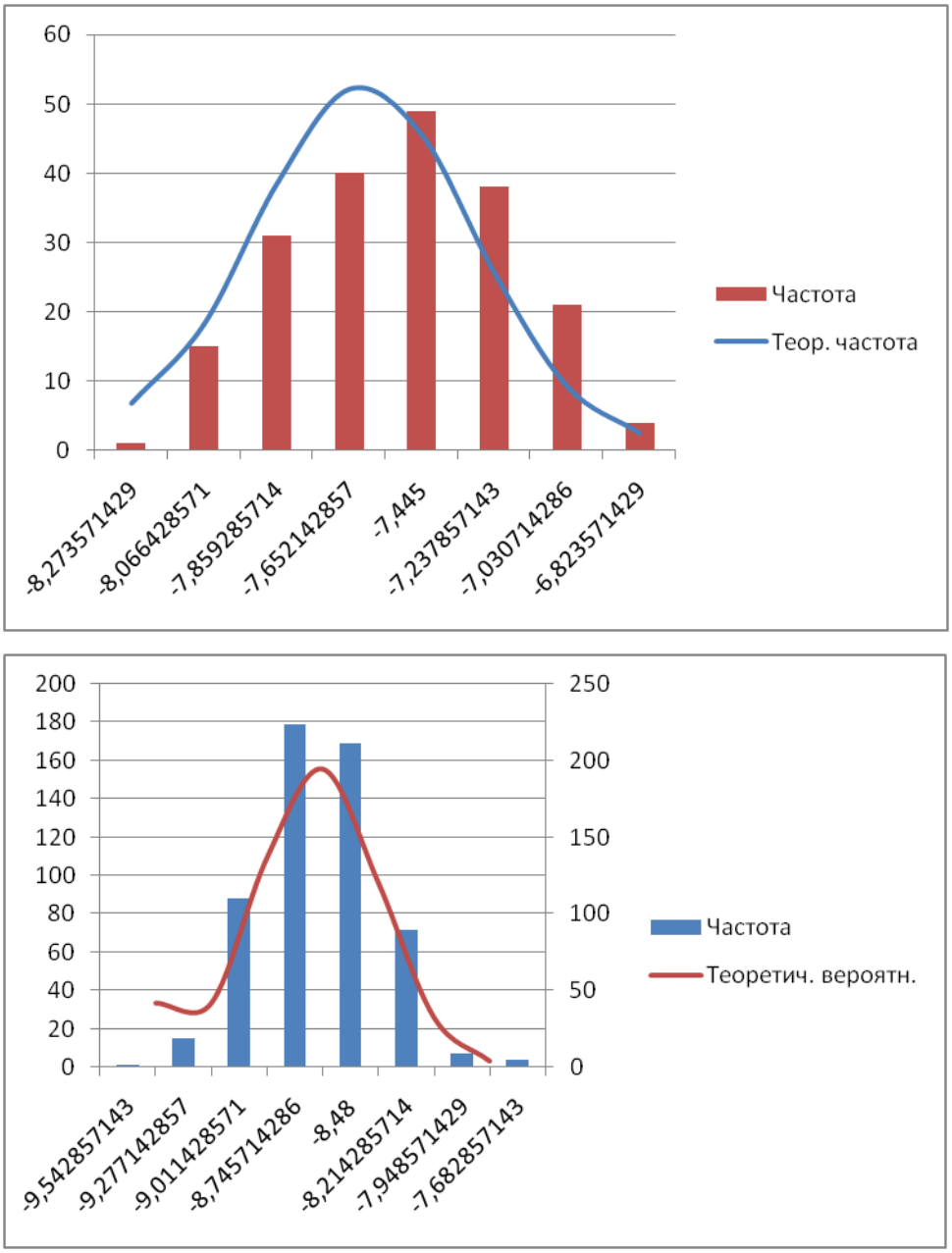


Рисунок 3.13 –Частотная гистограмма параметра Т76 ИС 1526ТЛ1. Вверху – партия №1 (199шт.), внизу – партия №2 (535 шт.)

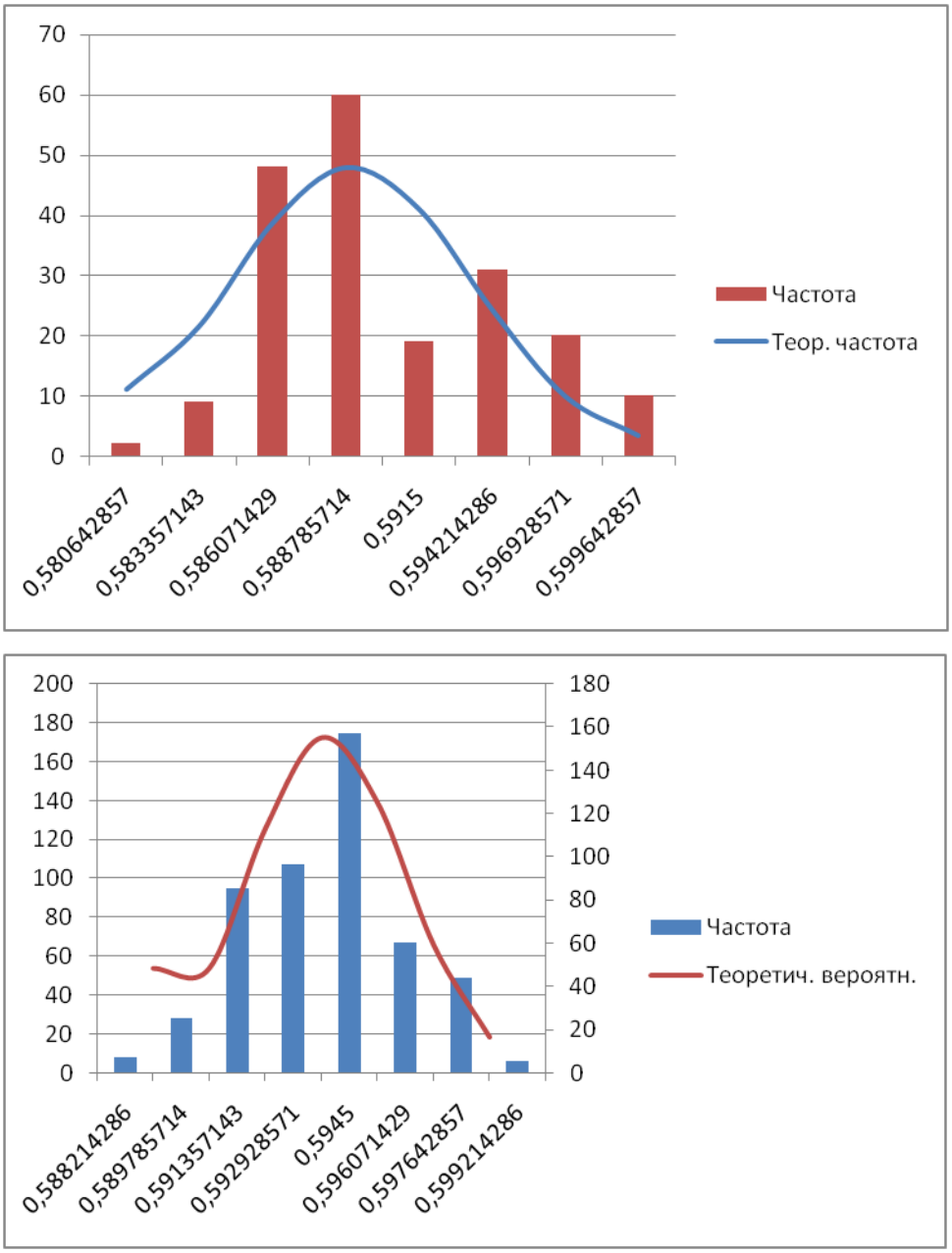


Рисунок 3.14 –Частотная гистограмма параметра Т89 ИС 1526ТЛ1. Вверху – партия №1 (199шт.), внизу – партия №2 (535 шт.)

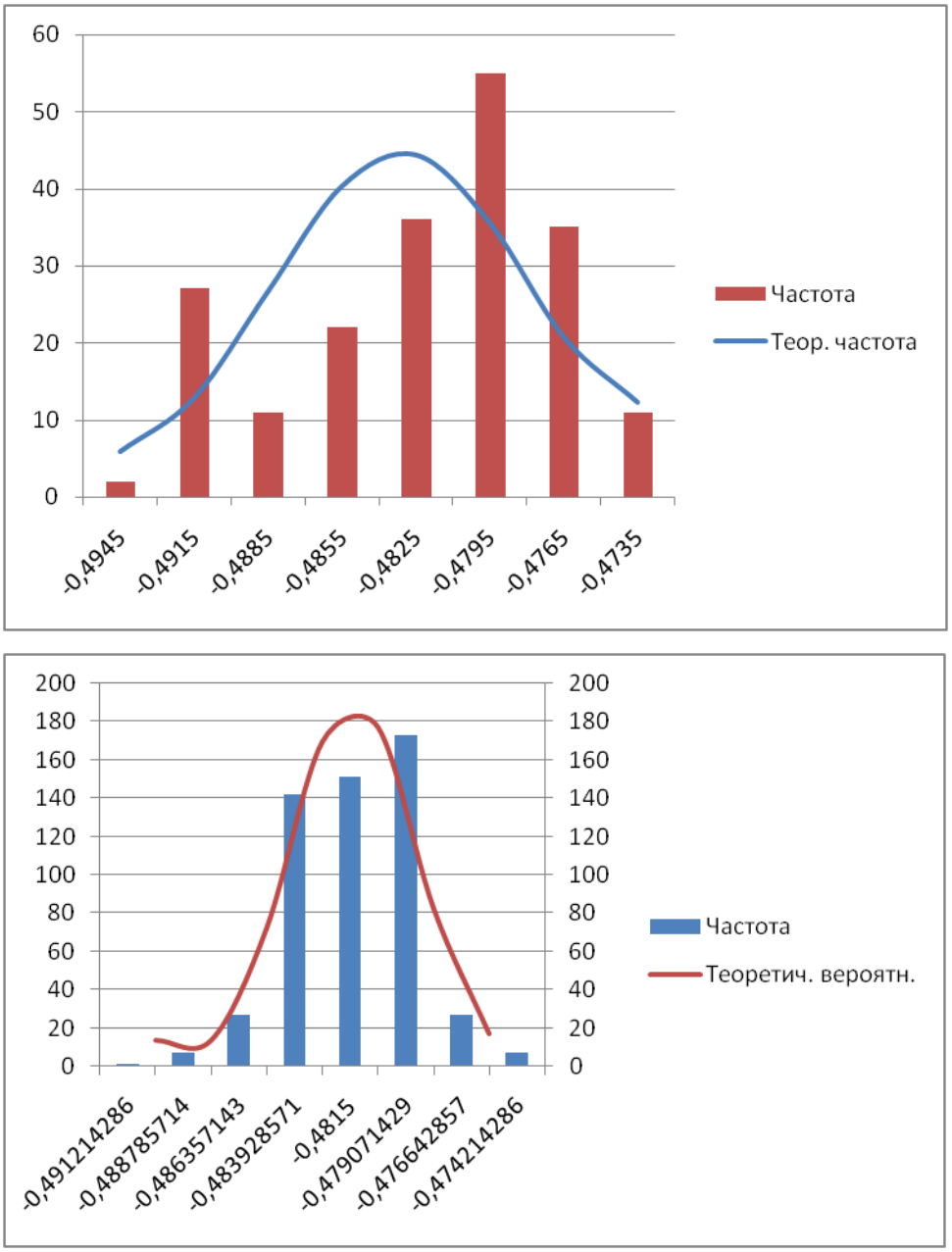


Рисунок 3.15 –Частотная гистограмма параметра Т110 ИС 1526ТЛ1. Вверху – партия №1 (199шт.), внизу – партия №2 (535 шт.)

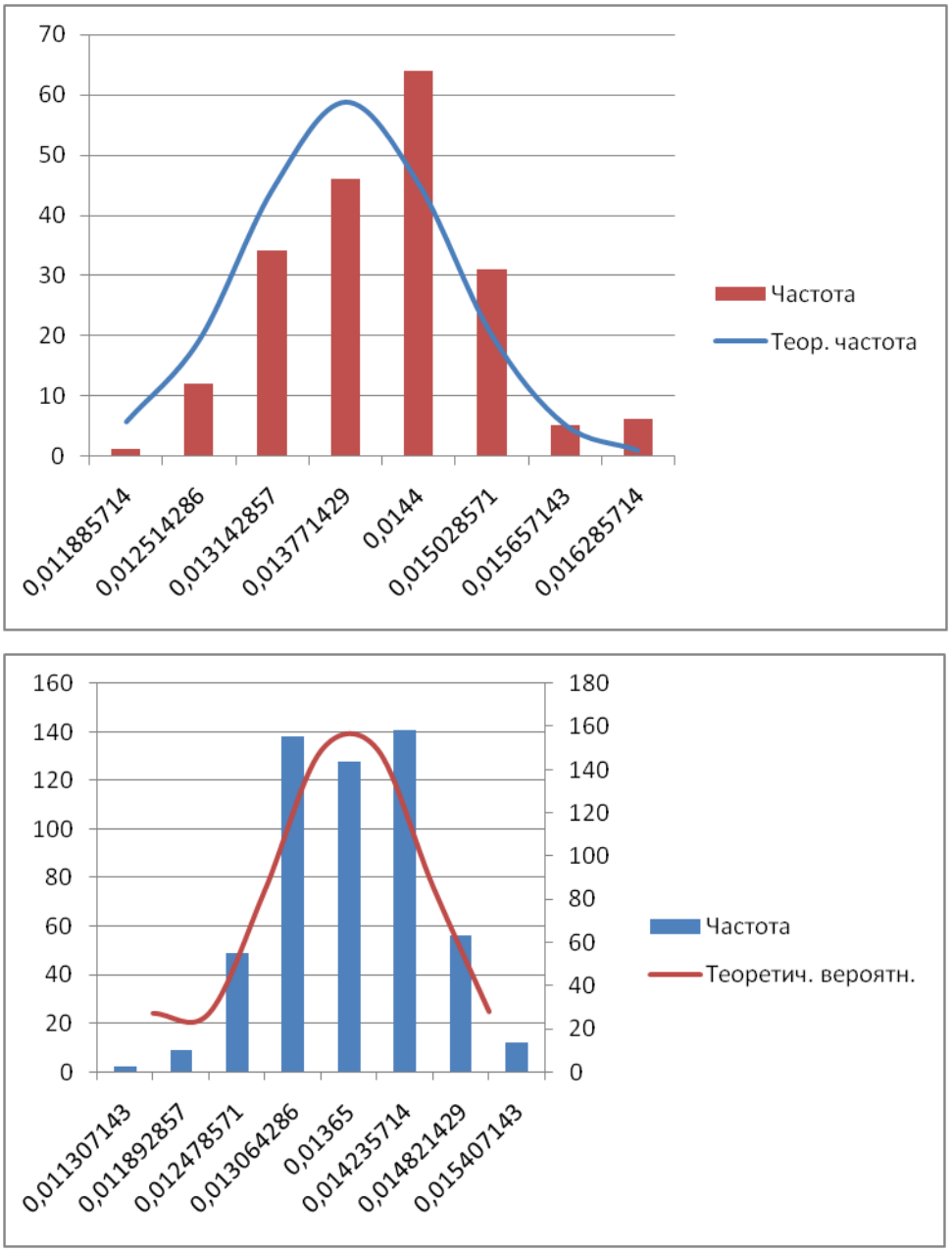


Рисунок 3.16 –Частотная гистограмма параметра Т143 ИС 1526ТЛ1. Вверху – партия №1 (199шт.), внизу – партия №2 (535 шт.)

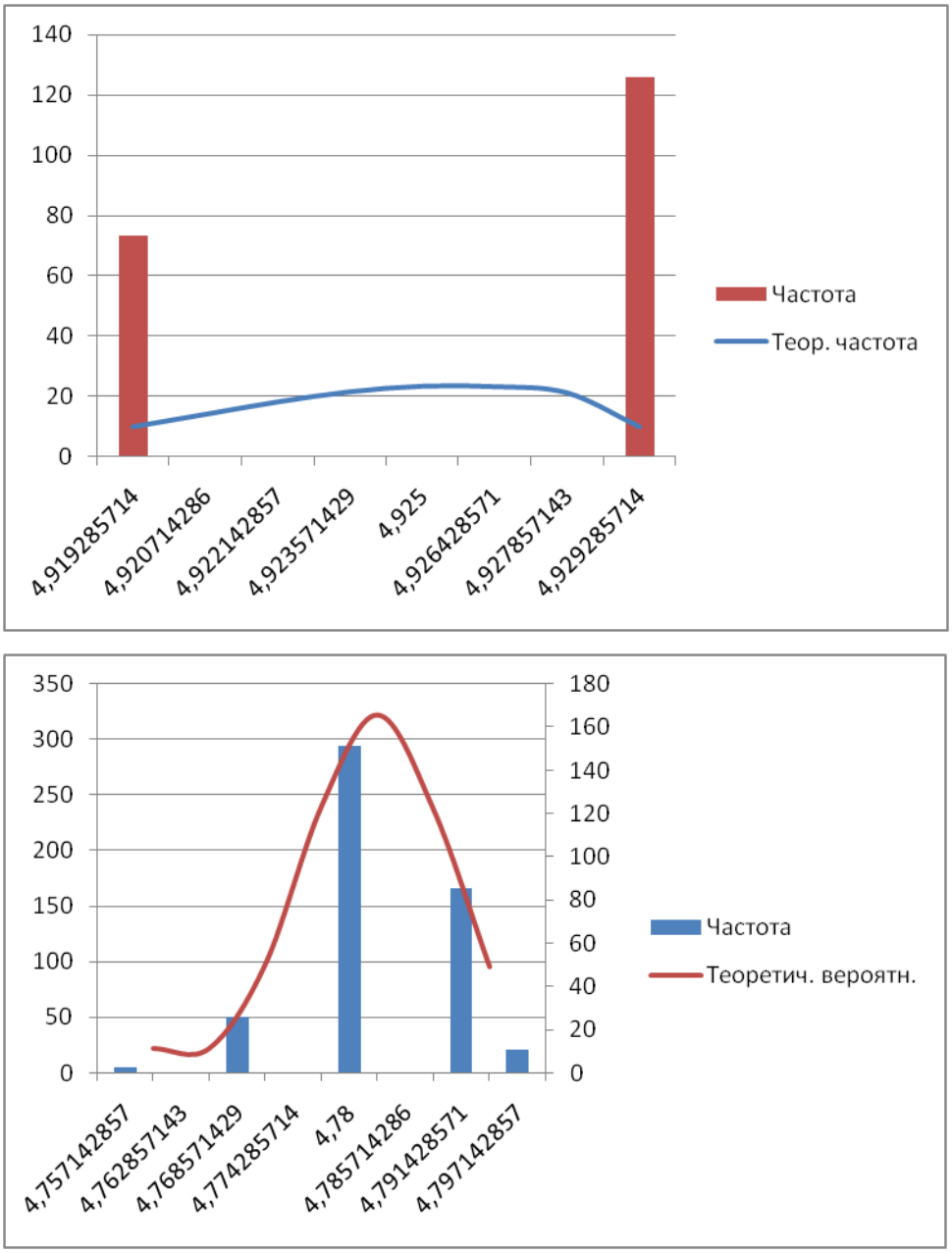


Рисунок 3.17 –Частотная гистограмма параметра Т154 ИС 1526ТЛ1. Вверху – партия №1 (199шт.), внизу – партия №2 (535 шт.)

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 [72]:

А) группа параметров, для которых нормальных характер распределения четко прослеживается. К таковым, например, для ИС 1526ТЛ1 можно отнести Т2-Т5 (рисунки 3.4-3.6), Т76-Т76 (рисунки 3.11-3.13), Т143 (рисунок 3.16).

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 в части фактической частоты попадания значений в интервалы. Отметим, что оценка правдоподобия нормальности распределения критерием Пирсона (критерием Хи-квадрат) [61] дает отрицательный ответ касательно гипотезы о нормальности распределения. К таким параметрам можно отнести, например, Т154 (рис. 3.17).

В) Параметры, для которых частотная гистограмма имеет вид, не соответствующий нормальному распределению (например, Т12, рисунок 3.8). В некоторых случаях можно сделать предположение о распределении, близком к экспоненциальному.

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%. Отметим, что именно для параметров группы А и параметров группы Б в основном установлены допустимые границы дрейфа параметров.

Отметим, что модель сборной партии электрорадиоизделий в виде смеси сферических или некоррелированных гауссовых распределений [68, 73, 175, 52] позволяет разделять составные части сборных партий, при этом процент ошибок достаточно велик.

Модель разделения смеси распределений на основе модели k-средних [32, 151, 31] также достаточно хорошо исследована, при этом показано, что нормирование исходных данных с использованием установленных допустимых границ дрейфа параметров [79] позволяет снизить процент ошибок при разделении сборной партии.

Отделив применением критерия эксцесса параметры, нормальность распределения которых не подтверждается, можно сформировать выборку, с параметры которой высокой вероятностью имеют нормальное распределение. К такой выборке можно применять модель разделения сборной партии электрорадиоизделий на основе разделения смеси гауссовых распределений. В этом случае можно сделать обоснованное предположение о том, что использование выборки, для параметров которой доказана нормальность распределения (например, с применением критерия эксцесса), позволит повысить адекватность модели и снизить процент ошибок.

Для проверки адекватности модели разделения сборной партии электрорадиоизделий мы в качестве тестового примера использовали две сборные партии: партия интегральных схем 1526ИЕ10 из 870 штук, составленная из четырех партий и сборная партия операционных усилителей 140УД25АВК из 53 штук [51].

Результаты представлены в таблицах 3.4 и 3.5.

Исследование характера вероятностных распределений отдельных параметров устройств электроники позволяет сделать вывод о том, что, по крайней мере, значительная часть этих параметров образуют многомерное гауссово распределение. Отсев параметров, имеющих распределение, отличное от нормального, позволяет повысить адекватность модели разделения сборных партий электронных устройств на основе смеси гауссовых распределений.

Таблица 3.4 - Сравнительные результаты разбиения сборной партии ИС 1526ИЕ10, сборная партия (870 шт.)

Модель	Партия №1		Партия №2		Партия №3		Партия №4		% ошибок
	Кол-во согласно модели	Ошибочно отнесенные к другим партиям	Кол-во согласно модели	Ошибочно отнесенные к другим партиям	Кол-во согласно модели	Ошибочно отнесенные к другим партиям	Кол-во согласно модели	Ошибочно отнесенные к другим партиям	
Фактически	146	-	229	-	424	-	71	-	-
Некореллир. гауссовы распределения по всем измерениям	244	25	196	41	358	78	72	1	16,7 %
Некореллир. гауссовы распределения с отбором измерений критерием эксцесса	125	26	221	8	414	10	110	0	5,05 %
Сферич. гауссовы распределения [90]	251	0	189	40	359	76	71	0	13,3 %

Таблица 3.5 - Сравнительные результаты разбиения сборной партии ИС 140УД25АВК, сборная партия (53 шт.)

Показатель	Значение
Фактически партия №1	30
Фактически партия №2	26
Всего	818
По результатам работы алгоритма классификации, построенном на модели k-средних [96]	
Партия №1	29
Партия №2	27
Ошибочные (шт./%)	3(5,3%)
По результатам работы алгоритма классификации, построенном на модели смеси сферических гауссовых распределений и нормированием по границам дрейфа параметров [90]	
Партия №1	29
Партия №2	27
Ошибочные (шт./%)	2 (3,57%)
По результатам работы алгоритма классификации, построенном на модели смеси некоррелированных гауссовых распределений и нормированием по границам дрейфа параметров	
Партия №1	29
Партия №2	27
Ошибочные (шт./%)	2 (3,57%)
По результатам работы алгоритма классификации, построенном на модели смеси некоррелированных гауссовых распределений и нормированием по границам дрейфа параметров с отбором измерений критерием эксцесса	
Партия №1	29
Партия №2	27
Ошибочные (шт./%)	2 (3,57%)

3.4 Критерии оценки адекватности моделей сборных партий электрорадиоизделий

Для выявления однородных партий используются алгоритмы кластеризации, некоторые подходы к этой проблеме рассмотрены в [151, 153, 32]. Одной из основных задач при этом является определение предполагаемого количества партий.

Задача определения числа групп.. Автоматическое определение числа групп является одной из труднейших задач в классификации данных. Большинство методов автоматического определения числа групп сводятся к проблеме выбора модели. Алгоритмы классификации запускаются при различном числе групп,

наилучшее значение выбирается на основании критерия компактности и взаимной удаленности выявленных групп в нормированном пространстве.

В литературе по автоматической группировке и кластерному анализу упоминаются следующие критерии: индекс Дэвиса-Боулдина (DBI) [114], индекс Калински-Харабаша [104], индекс Кржановски-Лая [159], критерий Хартигана [135], GAP-критерий [198], информационный критерий Байеса (BIC) [186], информационный критерий Акаике (AIC) [86], критерий силуэта [183].

Байесовский информационный критерий (Bayesian information criterion, BIC, иногда – Schwarz Criterion) - критерий выбора модели из класса параметризованных моделей, зависящих от разного числа параметров, используемый также в задачах автоматической группировки данных.

Критерий силуэта изначально был создан как средство интерпретации и подтверждения результатов автоматической группировки данных.

Значения различных критериев применительно к сборной партии интегральных схем космического применения [51] приведены в таблице 3.6.

Таблица 3.6 - Значения целевой функции и критериев оценки эффективности разбиения на группы в зависимости от их числа для изделия 1526ЛЕ5 [51]

k	Целевая функция	Критерий Хартигана	Критерий силуэта ускоренный	Критерий силуэта	Критерий BIC	Критерий AIC
1	84359	220,4895	-	-	334844	5302219
2	53968	206,9374	0,53918	0,41129	316973	3189458
3	37985	54,3987	0,57784	0,45594	295186	1730708
4	27329	46,9809	0,53645	0,38642	291719	1549824
5	21098	35,9868	0,48232	0,34088	287171	1347527
6	19865	23,0933	0,44891	0,31559	282926	1182413
7	18952	22,6444	0,43588	0,29702	281246	1113269
8	18091	20,4564	0,42419	0,28621	279008	1032668
9	17357	15,0933	0,39457	0,25063	277946	988834
10	17021	13,4691	0,38525	0,23961	276951	968357

Критерий силуэта пригоден к использованию только при условии наличия 2 или более кластеров в исследуемых данных. Следует провести оценку распределений частот и плотностей числовых характеристик на наличие многомодальности, что укажет на возможность разделения исходной выборки на группы, сходные по характеристикам объектов [37].

Проблема отсева «выбросов». Кроме надежного определения принадлежности каждого из векторов данных к тому или иному кластеру, критерием которой может служить служит силуэт каждого из векторов данных, требуется также определить, входит ли тот или иной вектор данных в какой-либо из кластеров, или же он является отдельным, далеко отстоящим от всех кластерных структур вектором (так называемым «выбросом» - “outlier”). Наличие «выбросов» - отдельно стоящих элементов вне кластеров – может быть вызвано различными причинами, например, ошибкой при проведении измерений каких-либо из параметров или же фактическим значительным отклонением этих параметров от значений, типичных для остальной части партии, у конкретного экземпляра ЭРИ вследствие брака или вследствие попадания единичных экземпляров из другой партии.

Такие «выбросы» в любом случае являются аномалией в рамках партии ЭРИ, и данные экземпляры не должны отбираться, например, для проведения разрушающего физического анализа. Так же и на них не могут быть распространены результаты РФА, выполненного на каких-либо из «типичных» экземпляров партии.

В настоящее время в литературе нет единого подхода к определению «выбросов», соответственно, нет и единой методологи.

Тем не менее, логично считать «выбросом» точку (вектор данных), которая отстоит на значительном расстоянии от других точек. Значительным будем считать расстояние, которое в q раз превосходит среднее внутрикластерное расстояние.

Для данной цели введем следующий критерий для каждого из векторов данных $A_1,\ldots,A_N$:

$$K_{outlier}(i) = \frac{\sum_{j \in C_i} \left\| A_i - A_j \right\| / |C_i|}{\sum_{k=1}^{N} \left(\sum_{j \in C_k} \left\| A_k - A_j \right\| / |C_k| \right)}.$$

Здесь C_i – кластер (множество векторов данных), которому принадлежит i-й

вектор данных, $|C_i|$ - его мощность (количество векторов данных в нем). Здесь для измерения расстояния используется та метрика или мера расстояния, которая используется и при решении задачи k-средних.

В числителе дроби – среднее расстояние от *i*-го вектора данных до других векторов «своего» кластера, в знаменателе – среднее внутрикластерное расстояние по всей партии ЭРИ.

Таким образом, например, для «отсева» (т.е. для причисления их к «выбросам») тех векторов данных, для которых расстояние до других векторов «своего» кластера в q раз превышает среднее внутрикластерное расстояние, требуется проверить условие $K_{outlier}>q$.

В нашем реализованном программном приложении используется модифицированный критерий:

$$K^*_{outlier}(i) = \frac{1}{1+K_{outlier}(i)}$$

и, соответственно, иное пороговое значение:

$$K^*_{outlier}(i) < q^*,$$

$$q^* = \frac{1}{1+q}.$$

Так, если требуется отсечь в качестве «выбросов» векторы данных, для которых среднее внутрикластерное расстояние превышено в два раза, следует установить значение

$$q^* = \frac{1}{1+2} \approx 0{,}33$$

Именно это значение мы использовали в наших экспериментах.

Таким образом, «выбросы» определяются проверкой условия

$$\frac{1}{1+\frac{\sum_{j\in C_i}\|A_i - A_j\|/|C_i|}{\sum_{k=1}^{N}\left(\sum_{j\in C_k}\|A_k - A_j\|/|C_k|\right)}} < q^*.$$

С целью проверки работоспособности метода автоматической группировки

электрорадиоизделий по производственным партиям мы исследовали партии интегральных схем, заведомо составленные из двух или трех различных производственных партий. Для визуального представления многомерных данных использовалось многомерное шкалирование [24]. На диаграмме многомерного шкалирования цифрами обозначены номера предполагаемых производственных партий, «X» и «?» - не классифицированные элементы. На графике критерия силуэта показана зависимость значения критерия от предполагаемого числа производственных партий.

На рис. 3.18 показано разбиение сборной партии операционных усилителей на две предполагаемые производственные партии, а также значения критерия силуэта для различных вариантов разбиения на разное число партий. Видно, что максимальное значение критерия силуэта четко указывает на число производственных партий, равное двум. Отметим, что разбиение всех элементов, кроме помеченных как ненадежно классифицированные, совпадает с их реальной принадлежностью к той или иной производственной партии. К ненадежно классифицированным отнесены элементы, лежащие за пределами какого-либо из кластеров.

При этом наиболее важным вопросом остается адекватность классификации. Были проведены эксперименты на экзаменационных выборках ИС, состав элементов с разбивкой на производственные партии был заранее известен. В таблице 3.7 показаны результаты разбиения этих выборок.

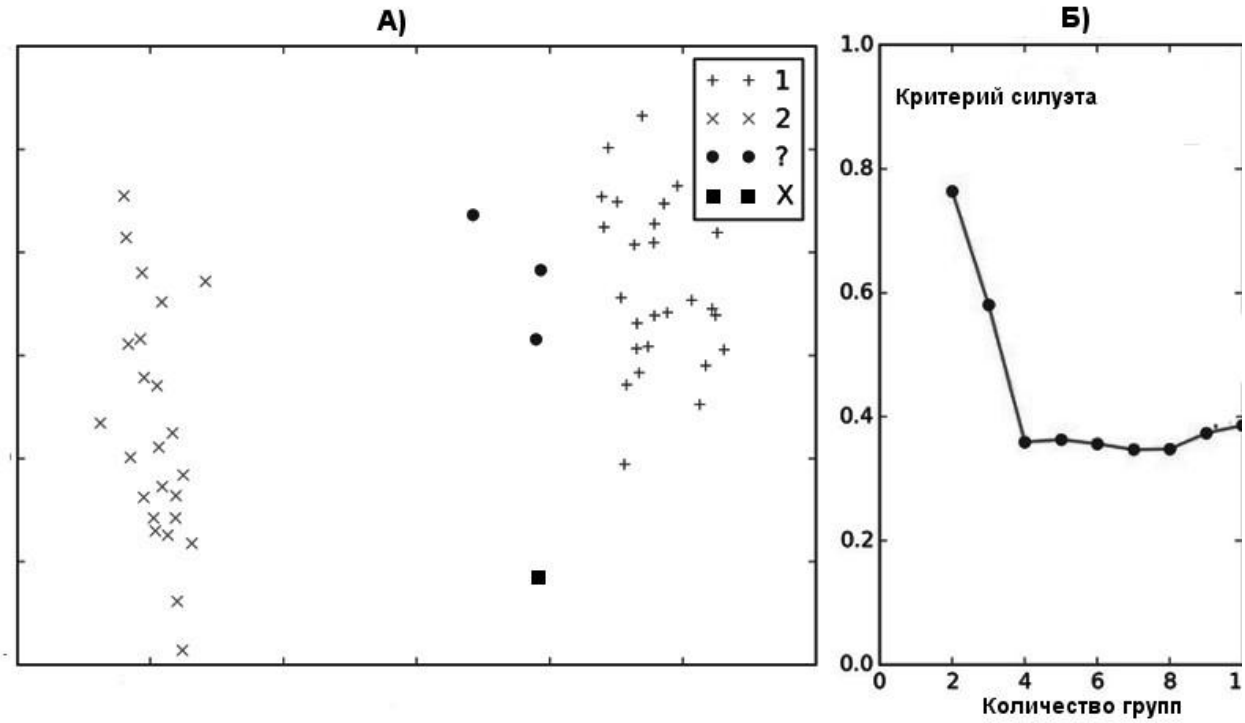


Рисунок 3.18– Сборная партия ИС 140УД17АВК, разбиение на группы. А) – многомерноешкалирование; Б) – критерий силуэта

Таблица 3.7 - Результаты разбиения экзаменационных выборок интегральных схем

Показатель	ИС 140УД25АС1ВК	ИС 140УД17АВК
Партия №1	30	26
Партия №2	26	24
Всего	56	50
По результатам работы алгоритма классификации		
Партия №1	27	26
Партия №2	26	24
Неклассифицированные	3 (8,9%)	0 (0%)
Ошибочные	0	0

Таким образом, применение критерия силуэта позволяет эффективно определять число производственных партий в сборной партии, что, в свою очередь, позволяет повысить эффективность алгоритма автоматической группировки и более точно классифицировать электрорадиоизделия.

3.5 Общая схема принятия решений по приемке партий электрорадиоизделий

В рамках диссертационной работы была обновлена схема метода принятия решений по комплектации электронных узлов космических аппаратов [27].

На рис. 3.19 приведена обновленная BPMN-схема принятия решения о приемке спецпартии. Схема, предложенная в [27] обновлена в части этапа выделения однородных партий методом разделения смеси распределений.

Для повышения надежности программных модулей систем управления, участвующих в принятии решения, целесообразно использовать методы повышения отказоустойчивости [71]. В частности, мы дополняем действующую систему автоматической группировки электрорадиоизделий по однородным производственным партиям на основе модели k-средних программной подсистемой на основе модели разделения смеси распределений. Такой подход позволяет обеспечить лицо, принимающее решения о приемке партий и о направлении их экземпляров на разрушающий физический анализ, дополнительной информацией. Одна модель кластеризации позволяет верифицировать результаты другой модели, а при несовпадении результатов в условиях повышенных требований к качеству, предъявляемых космической промышленностью, предлагается отказаться от использования спорных экземпляры изделий при комплектовании бортовой аппаратуры.

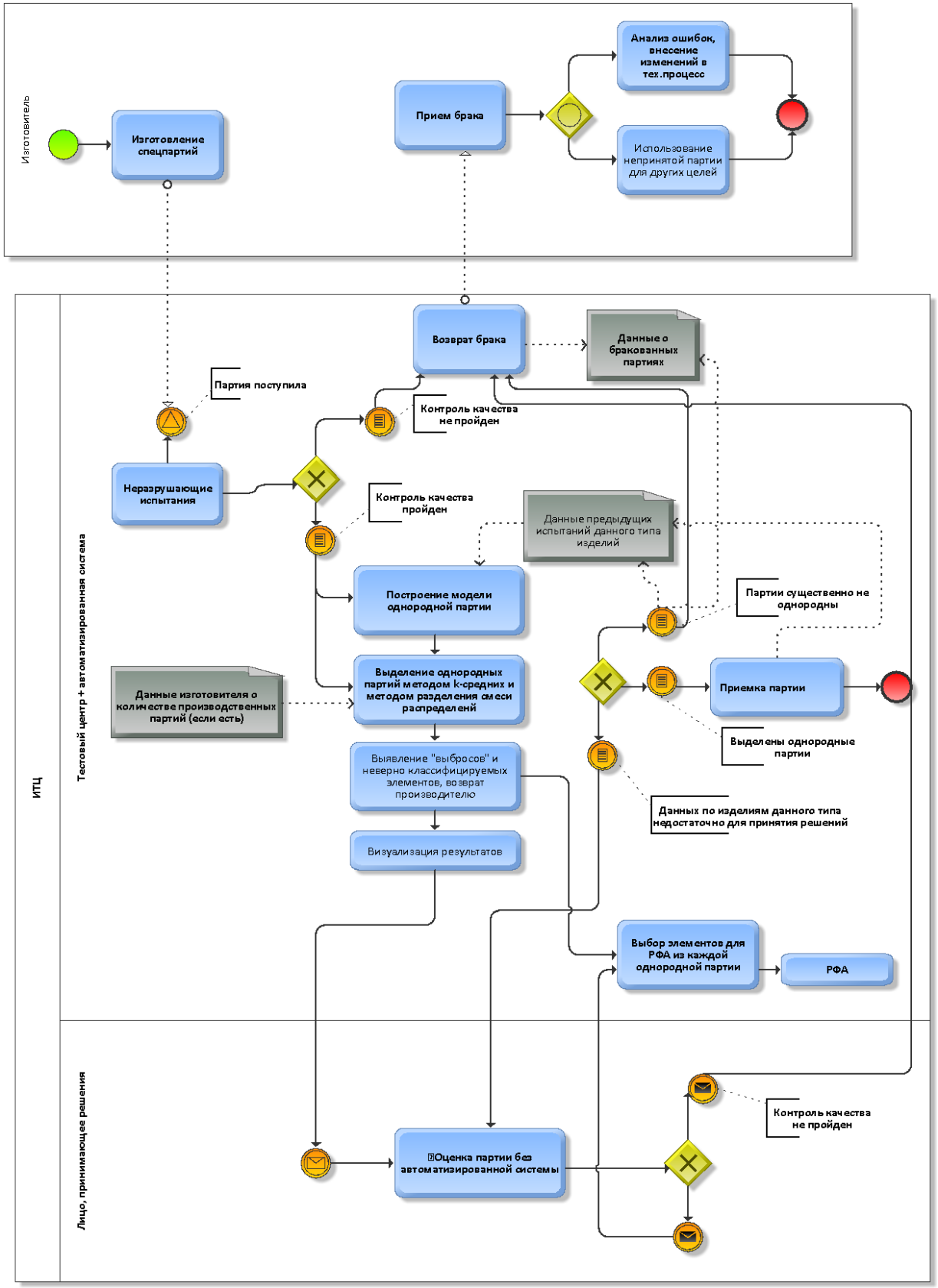


Рисунок 3.19 – Обновленная BPMN-схема процесса формирования спецпартии электронной компонентной космического применения (ЭКБ)

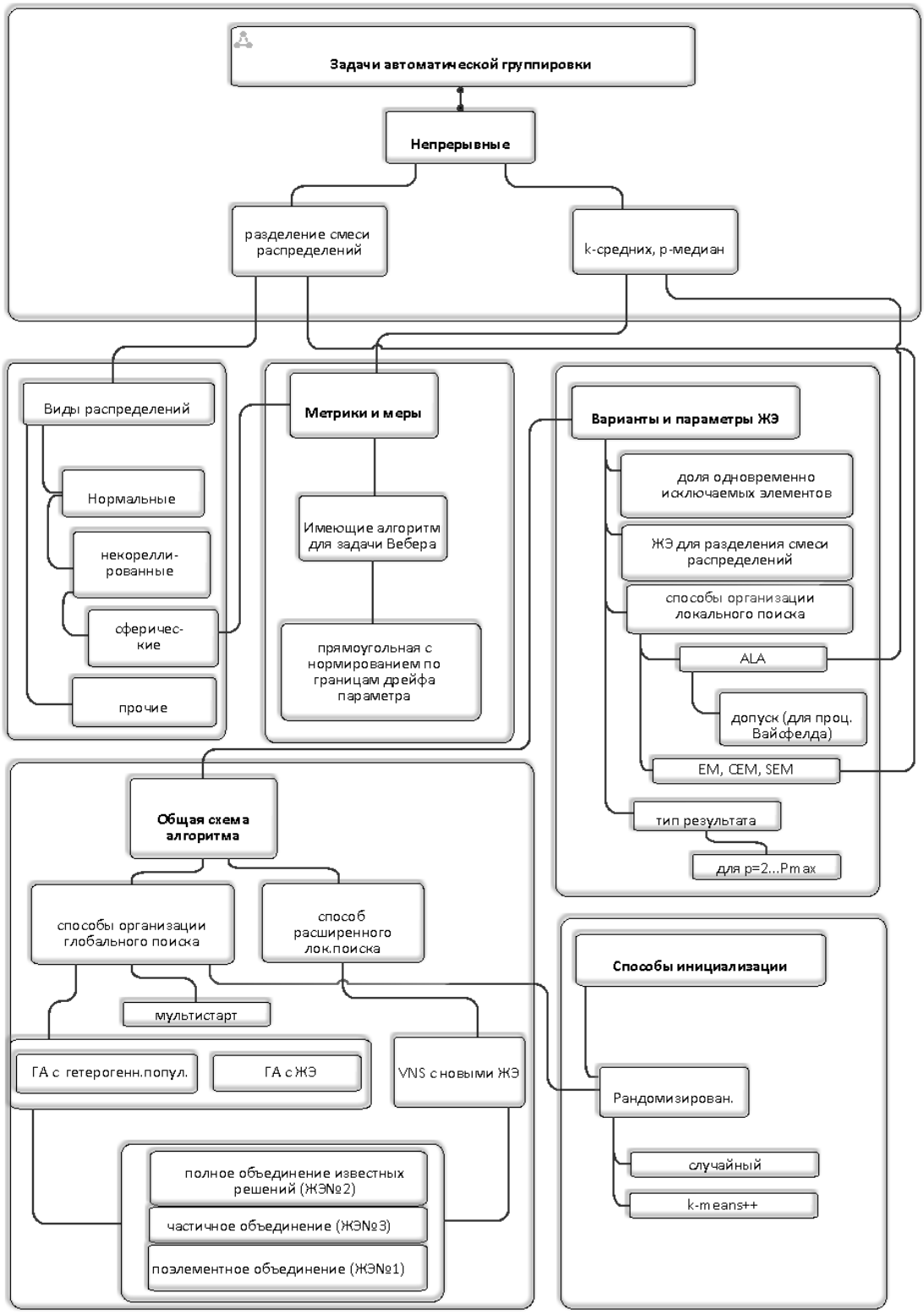


Рисунок 3.20 – Концептуальная схема системы разделения сборных партий электрорадиоизделий (ЭРИ) космического применения по результатам тестовых испытаний

На схеме (рис. 3.20) показаны задачи, модели, алгоритмы и их возможные взаимосвязи, которые могут быть задействованы при построении эффективной системы автоматической группировки электрорадиоизделий по однородным производственным партиям.

Результаты Главы 3

Построена новая модель разделения сборных партий промышленной продукции на однородные партии на основе модели разделения смеси распределений. Таким образом, на примере задачи выявления однородных производственных партий ЭРИ из сборной партии по данным неразрушающих тестовых испытаний показано, что новая модель позволяет вдвое снизить процент ошибок разделения сборной партии. 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, позволяющая 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, т.е. повысить точность кластеризации.

СПИСОК ЛИТЕРАТУРЫ

1. Агеев, А.А. Полиномиальный алгоритм решения задачи размещения на цепи с одинаковыми производственными мощностями предприятий / А.А. Агеев, Э.Х. Гимади, А.А. Курочкин // Дискретный анализ и исследование операций.- 2009.- Т.16, № 5.- С. 3–18.
2. Айвазян С. А., Бежаева.И, Староверов О. В., Классификация многомерных наблюдений. М., Статистика, 1974. 240 с.
3. Алексеев, О.Г.Некоторые алгоритмы решения задачи о покрытии и их экспериментальная проверка на ЭВМ / О.Г. Алексеев, В.Ф.Григорьев // Журнал вычислительной математики и математической физики. - 1984. - Т.24, №10.- С. 1565-1570.
4. Алексеева, Е.В. Алгоритмы локального поиска для задачи о р-медиане с предпочтениями клиентов: дис. ... канд. физ.-мат. наук.- Новосибирск: Институт математики им. С.Л.Соболева СО РАН.- 2007.-С. 92.
5. Андреев, В.Л. Классификационные построения в экологии и систематике / В.Л. Андреев.- М.:Наука.- 1980.- С.142.
6. Антамошкин, А.Н. Алгоритм случайного поиска для обобщенной задачи Вебера в дискретных координатах / А.Н. Антамошкин, Л.А. Казаковцев // Информатика и системы управления.- 2013.- № 1(35).- С. 87-98.
7. Арлазаров, В.В. Структурный анализ текстовых полей в системах потокового ввода оцифрованных документов /В.В. Арлазаров, В.М. Кляцкин, О.А. Славин // Труды ИСА РАН.- 2015.- Т. 65, вып. 1.- С.75-81.
8. Барахнин, В.Б. Кластеризация текстовых документов на основе составных ключевых термов / В.Б. Барахнин, Д.А.Ткачев// Вестник НГУ. Серия: Информационные технологии.- 2010.- Т.8, вып.2.- С. 5-14.
9. Барахнин, В.Б. О задании меры сходства лоя кластеризации текстовых документов / В.Б. Барахнин, В.А. Нехаева, А.М. Федотов // Вестник НГУ. Серия: Информационные технологии.- 2008.- Т.6, вып.1.- С.3-9.

10. Береснев, В.Л. Экстремальные задачи стандартизации /В.Л. Береснев, Э.Х. Гимади, В.Т. Дементьев.- Новосибирск: Наука, 1978.- 333 С.
11. Бериков, В.Б. Современные тенденции в кластерном анализе / В.Б.Бериков, Г.С. Лбов // Всероссийский конкурсный отбор обзорно-аналитических статей по приоритетному направлению "Информационно-телекоммуникационные системы". Новосибирск: Институт математики им. С.Л. Соболева СО РАН.- 2008.- 26 с. [Электронный ресурс] Режим доступа URL http://www.ict.edu.ru/ft/005638/62315e1-st02.pdf (дата обращения 12.02.2016).
12. Борисенко, В.И. Сегментация изображения (состояние проблемы) / В.И.Борисенко, А.А.Златопольский, И.Б. Мучник// Автомат. и телемех.- 1987.- вып. 7.- С. 3-56.
13. Браверман, Э.М. Структурные методы в обработке эмпирических данных / Э.М. Браверман, И.Б. Мучник.- М.: Наука.- 1983.- С.464.
14. Бродский Б. Е. , Дарховский Б. С., Проблемы и методы вероятностной диагностики, Автомат. и телемех., 1999, выпуск 8, 3–50
15. Васильев, И.Л. Новые нижние оценки для задачи размещения с предпочтениями клиентов / И.Л. Васильев, К.Б. Климентова, Ю.А. Кочетов // Журнал вычислительной математики и математической физики.- 2009.- т.49, вып. 6.- С.1055-1066.
16. Галушкин А. И., Кай мин В, А., Моментный подход к решению задачи самообучения системы распознавания образов. Тр. Моск. ин-та электрон. машиностр., 1971, вып. 14, 139—146
17. Гимади, Э. Х. Задача стандартизации с данными произвольного знака и связными, квазивыпуклыми и почти квазивыпуклыми матрицами / Э.Х. Гимади // Управляемые системы. Сб. науч. тр. Вып. 27. — Новосибирск: Ин-т математики СО АН СССР.- 1987.- С. 3-11.

18. Гимади, Э.Х. Эффективные алгоритмы для решения многоэтапной задачи размещения на цепи / Э.Х.Гимади // Дискретн. анализ и исслед. Опер..- 1995.- том 2, № 4.- С. 13–31.
19. Граничин, О. Н. Рандомизированные алгоритмы оценивания и оптимизации при почти произвольных помехах / О.Н. Граничин, Б.Т. Поляк.- М.: Наука. -2003.-С. 291.
20. Дорофеюк А. А., Алгоритмы обучешя машины раепоздавамию образов без учителя, основанные на методе потенциальных функций. Автоматика и телемеханика, 1966, № 10, 78—87
21. Епанечников, В.А. Непараметрическая оценка многомерной плотности вероятности / В.А. Епанечников // ТВиП.- 1969.- Т.14.- С.156-161.
22. Еремеев, А.В. Генетический алгоритм с турнирной селекцией как метод локального поиска / А.В. Еремеев // Дискретный анализ и исследование операций.- 2012.- Т. 19, № 2 .- С. 41-53.
23. Загоруйко, Н.Г. Прикладные методы анализа данных и знаний / Н.Г. Загоруйко.- Новосибирск: ИМ СО РАН.- 1999. С. 270.
24. Зиновьев А. Ю., Визуализация многомерных данных, Красноярск, Изд. КГТУ, 2000.
25. Иванова, Н.В. Определение параметров сглаживания в непараметрических оценках функции плотности по выборке / Н.В. Иванова, К.Т. Протасов // Математическая статистика и ее приложения. Томск: изд-во Томск, гос. Ун-та.- 1982.- Вып. 8.- С. 50-65.
26. Исаенко О. К. , Урбах В.Ю. Разделение смесей распределений вероятностей на их составляющие, Итоги науки и техн. Сер. Теор. вероятн. Мат. стат. Теор. кибернет., 1976, том 13, 37–58
27. Казаковцев Л. А. Метод жадных эвристик для систем автоматической группировки объектов: Дисс… докт. Техн. наук. – Красноярск, 2016. – 429 с.

28. Казаковцев Л. А., Орлов В. И., Ступина А. А. Выбор метрики для системы автоматической классификации электрорадиоизделий по производственным партиям // Программные продукты и системы. 2015. № 2. С. 124–129. Doi: 10.15827/0236-235X.110.124-129.
29. Казаковцев Л.А. Модификация генетического алгоритма с жадной эвристикой для непрерывных задач размещения и классификации / Л.А. Казаковцев, А.А. Ступина, В.И. Орлов // Системы управления и информационные технологии.- 2014.- вып.2(56).- С.35-39.
30. Казаковцев Л.А., Антамошкин А.Н. Метод жадных эвристик для задач размещения // Вестник СибГАУ. 2015. №2. С.317-325.
31. Казаковцев Л.А., Масич И.С., Орлов В.И., Федосов В.В. Быстрый детерминированный алгоритм для классификации электронной компонентной базы по критерию равнонадежности // Системы управления и информационные технологии. 2015. Т. 62. № 4. С. 39-44.
32. Казаковцев Л.А., Орлов В.И., Ступина А.А. Выбор метрики для системы автоматической классификации электрорадиоизделий по производственным партиям // Программные продукты и системы. 2015. No 2. С. 124129. doi: 10.15827/0236235X.110.124129.
33. Казаковцев, Л.А. Алгоритм случайного поиска для обобщенной задачи Вебера в дискретных координатах / А.Н.Антамошкин, Л.А.Казаковцев// Информатика и системы управления. – 2013. –Вып. 1. –С. 87-98.
34. Казаковцев, Л.А. Модификация генетического алгоритма с жадной эвристикой для непрерывных задач размещения и классификации / Л.А. Казаковцев, А.А. Ступина, В.И. Орлов // Системы управления и информационные технологии.- 2014.- вып.2(56).- С.35-39.
35. Казаковцев, Л.А. Параллельный алгоритм для p-медианной задачи / Л.А.Казаковцев, А.Н. Антамошкин, М.Н. Гудыма // Системы управления и информационные технологии.- 2013.- № 52 (2.1).- 2013.- Р.124–128.

36. Кононов В.К., Малинин В.Г., Оспищев Д.А., Попов В.Д. Отбраковка потенциально-ненадежных интегральных микросхем с использованием радиационно-стимулирующего метода / В сб. Радиационно-надежностные характеристики изделий электронной техники в экстремальных условиях эксплуатации. Под редакцией Ю.Н. Торгашова // СПб.: Издательство РНИИ «Электронстандарт», 1994, 96с.
37. Коплярова Н. В., Орлов В. И. Об исследовании компьютерной системы диагностики электрорадиоизделий на основе данных испытаний // Вестник СибГАУ. 2014. Вып. 1(53), С. 24–30.
38. Коплярова, Н.В. О непараметрических моделях в задаче диагностики электрорадиоизделий / Н.В.Коплярова, В.И.Орлов, Н.А.Сергеева, В.В.Федосов // Заводская лаборатория: диагностика материалов.- 2014.- №80(7).- С.37-77.
39. Королев В. Ю. ЕМ-алгоритм, его модификации и их применение к задаче разделения смесей вероятностных распределений. Теоретический обзор // ИПИ РАН. М., 2007. С. 94.
40. Космический аппарат "Sesat" со сроком активного существования 10 лет. Принципы, методы и результатыкомплектации аппаратуры электрорадиоизделиями: Технический отчет / Козлов А.Г., Исляев Ш.Н., ФедосовВ.В., Коновалов В.И., Белов С.А., Куклин В.И., Орлов В.И., НПО прикладной механики, Красноярск 26, 1999.-408с.
41. Кочетов, Ю.А. Методы локального поиска для дискретных задач размещения: дис. … доктора физ.-мат. Наук: 05.13.18: защищена 19.01.2010.- Новосибирск: Институт математики им.Соболева.- 2010.- С.259.
42. Кочетов Ю. А. , Младенович Н., Хансен П., “Локальный поиск с чередующимися окрестностями”, Дискретн. анализ и исслед. опер., сер. 2, 10:1 (2003), 11–43

43. Куклин В.И, Орлов В.И., Федосов В.В. Результаты работ по обеспечению качества электрорадиоизделий отечественного производства для комплектования бортовой аппаратуры космических аппаратов за период 01.2008г.-06.2009г. // VIII Российская научно-техническая конференция. Электронная компонентная база космических систем. М.: 2009. С. 64–66.
44. Малышев М.М., Малинин В.Г., Куликов И.В., Торгашов Ю.Н, Ужегов М.В. Методология оценки радиационной надежности ИЭТ в условиях низкоинтенсивных ионизирующих излучений. / В сб. Радиационно-надежностные характеристики изделий электронной техники в экстремальных условиях эксплуатации. Под редакцией Ю.Н. Торгашова // СПб.: Издательство РНИИ «Электронстандарт», 1994, 96с.
45. Мальцева Н. И., Оценка параметров смеси мер по методу моментов. Сб. аспирантск. работ. Казаиск. ун-т. Точн. науки Мат. Мех. Физ., Казань, 1969.
46. Медведев, А.В. Непараметрические оценки плотности вероятности и ее производных / А.В. Медведев // Автоматизация промышленного эксперимента.- Фрунзе: Илим.- 1973.- с. 22-31.
47. Миленький А. В., Определение статистических характеристик распознаваемых образов в режиме самообучения. Кибернетика, 1967, № 3, 56—61.
48. Модель околоземного космического пространства: В 3-х т. Под ред. академика Вернова С.Н. Издание седьмое // М.: МГУ, 1983 г.
49. Мырова Л.О., Чепиженко А.З. Обеспечение стойкости аппаратуры связи к ионизирующим и электромагнитным излучениям. 2-е изд., перераб. и доп. // М.: Радио и связь, 1988, 296 с.
50. Надарая Э.А. Об оценке регрессии / Э.А. Надарая // ТВиП.- 1964.- Т. 9, №1.- С. 157-159.
51. Орлов В.И., Казаковцев Л.А., Масич И.С. Применение критерия силуэта в алгоритме автоматической группировки электрорадиоизделий космического применения // Вестник Сибирского государственного

аэрокосмического университета им. академика М.Ф. Решетнева. 2016. Т. 17. № 4. С. 883-890.

52. Орлов В.И., Сташков Д.В., Гудыма М.Н., Казаковцев Л.А. Применение EM-алгоритма к задаче автоматической группировки электрорадиоизделий // Решетневские чтения: Материалы XX Юбилейной международной научно-практической конференции (9-12 ноября 2016). 2016. Т.2. С. 72-73.
53. Пакет «EMCluster». URL https://cran.r-project.org/web/packages/EMCluster/index.html
54. Пакет «mclust». URL https://cran.r-project.org/web/packages/mclust/index.html
55. Перечень ЦК-1/96. Изделия электронной техники, допускаемые для применения в аппаратуре космического аппарата «Ямал» с 10-летним сроком активного существования // АО ИТЦ «Циклон», 1997, 90 с.
56. Пиз Р.Л, Джонстон А.Х., Азаревич Дж.Л. Радиационные испытания полупроводниковых приборов для космической электроники // ТИИЭР, 1988, Т.76, № 11, стр. 126-145
57. Радиационная стойкость бортовой аппаратуры и элементов космических аппаратов // I Всесоюзная научно-техническая конференция. Материалы конференции. Томск, 1991, 257с.
58. Радиационная стойкость материалов радиотехнических конструкций: Справочник. Под ред. Н.А Сидорова, В.К. Князева // М.: Советское радио, 1976, 567 с.
59. Решение № SST-TP-97006 о квалификации электрорадиоизделий на соответствие требованиям космического аппарата с 10-летним сроком активного существования (Редакция 1-97) // АО ИТЦ «Циклон», 1997, 108 с.
60. Рубан А.И. Идентификация и чувствительность сложных систем / А.И. Рубан.- Томск:Изд-во Томск, гос. Ун-та.- 1982.- С.302.
61. Рубан А.И. Методы анализа данных, Красноярск: ИПЦ КГТУ, 2004. 319 с.

62. Семенкин, Е.С. Коэволюционный генетический алгоритм решения сложных задач условной оптимизации / Е.С. Семенкин, Р.Б. Сергиенко // Вестник СибГАУ.- 2009.- №2.- С. 17-21.
63. Семѐнкин, Е.С. Эволюционные методы моделирования и оптимизации сложных систем: Конспект лекций /Е.С. Семѐнкин, М.Н. Жукова, В.Г. Жуков, И.А. Панфилов, В.В. Тынченко.- Красноярск: СФУ.- 2007.- 515 С.
64. Сергиенко, И.В. Метематические модели и методы решения задач целочисленной оптимизации / И.В. Сергиенко, 2-е изд.,доп. и перераб.- Киев: Наукова думка.- 1988.- С.472.
65. Славин, О.А. Алгоритмы распознавания шрифтов в печатных документах / О.А. Славин // Информационные технологии и вычислительные системы.- 2010.- №3.- С. 27-38.
66. Справочник по прикладной статистике / Под ред. Э. Ллойда, У. Ледермана, С.А. Айвазяна и др.- М.: Финансы и статистика.- 1989. - Т.1. - С.510.
67. Среда моделирования R. URL https://cran.r-project.org/
68. Сташков Д.В. Алгоритмы поиска с чередующимися окрестностями для задачи разделения смеси распределений // Системы управления и информационные технологии. 2017. №1.
69. Сташков Д.В. Выявление однородных партий изделий космической радиоэлектроники на основе разделения смеси сферических гауссовых распределений В. И. Орлов, Л. А. Казаковцев, И. Р. Насыров, А. Н. Антамошкин// Вестник СибГАУ. – 2017.Т.18. – вып. 1. – С. 69-77.
70. Сташков Д.В. Генетические алгоритмы метода жадных эвристик для серии задач разделения смеси распределений/ Гудыма М.Н. // Информационные технологии моделирования и управления. — 2017. — №3(105) С. 181-191.
71. Сташков Д.В. Задача формирования отказоустойчивого программного комплекса управления промышленным объектом /Сташков Д.В., Казаковцев Л.А., Насыров И.Р. // Фундаментальные исследования.— 2015.— вып. 5.1 — С.149-153.

72. Сташков Д.В. Моделирование сборной партии электрорадиоизделий в виде смеси вероятностных распределений/ Насыров И.Р., Попов А.М., Орлов В.И. // Экономика и менеджмент систем управления. — 2017. — №2.2(24) С. 282-289.
73. Сташков Д.В., Орлов В.И., Насыров И.Р., Казаковцев Л.А. Применение ЕМ-алгоритма со сферическим гауссовым распределением к задаче классификации промышленной продукции // Экономика и менеджмент систем управления, 2017, Т.23, №1.1, С.185-193.
74. Стойкость изделий электронной техники к воздействию факторов космического пространства и электрических импульсных перегрузок: Справочник. Т. XII. 4-е изд. Книга 2. Термовакуумные и электрические воздействия // ВНИИ «Электронстандарт», 1990, 162 с.
75. Субботин, В., Стешенко В. Проблемы обеспечения бортовой космической аппаратуры космических аппаратов электронной компонентной базой // Компоненты и технологии. – 2011. – вып.11. – с.10-12
76. Федосов В. В., Казаковцев Л. А., Гудыма М. Н. Задача нормировки исходных данных испытаний электрорадиоизделий космического применения для алгоритма автоматической группировки // Информационные технологии моделирования и управления.2016. № 4. С. 263–268.
77. Федосов В.В. Вопросы обеспечения работоспособности электронной компонентной базы в аппаратуре космических аппаратов: учеб.пособие. Сиб. гос. аэрокосмич. Ун-т. – Красноярск, 2015. – 68 С.
78. Федосов В.В., Казаковцев Л. А., Гудыма М.Н. Задача нормировки исходных данных испытаний электрорадиоизделий космического применения для алгоритма автоматической группировки // Информационные технологии моделирования и управления. 2016. No 4. С. 263-268.
79. Федосов В.В., Казаковцев Л.А., Масич И.С. Метод нормировки исходных данных испытаний электрорадиоизделий космического применения для

алгоритма автоматической группировки // Системы управления и информационные технологии. 2016. Т. 65. № 3. С. 92-96.

80. Федосов В.В., Орлов В.И. Минимально необходимый объем испытаний изделий микроэлектроники на этапе входного контроля // Известия высших учебных заведений. Приборостроение. 2011. т.54, № 4. стр. 58-62.
81. Харченко В.С., Юрченко Ю.Б. Анализ структур отказоустойчивых бортовых комплексов при использовании электронных компонентов Industry // Технология и конструирование в электронной аппаратуре, 2003. №2. С. 3–10.
82. Черезов Д. С., Тюкачев Н. А. Обзор основных методов классификации и кластеризации данных // Вестник Воронеж. гос. ун-та. Сер. «Системный анализ и информационные технологии». 2009. Вып. 2.
83. Шенк Х. Теория инженерного эксперимента. Пер. с англ. Е. Г. Коваленко под ред. Н. П. Бусленко // М.: Мир, 1972, 382 с.
84. Шлезингер М. И., Взаимосвязь обучения и самообучения в распознавании образов. Кибернетика, 1968, № 2, 81—88.
85. Agarwal, C.C. Optimized crossover for the independent set problem/ C.C.Agarwal, J.B.Orlin, R.P. Tai // Operations research.- 1997.- Vol. 45, issue 2.- P. 226 - 234.
86. Akaike, H. A new look at the statistical model identification// IEEE Transactions on Automatic Control, 1974,Vol.19 (6),P.716–723. doi:10.1109/TAC.1974.1100705
87. Akhmedova, Sh.A. New optimization metaheuristic based on co-operation of biology related algorithms / Sh.A. Akhmedova, E.S. Semenkin // Vestnik SibGAU.-No. 4 (50).-2013.- P. 92-99.
88. Alp, O. An Efficient Genetic Algorithm for the p-Median Problem / O. Alp, E. Erkut, Z. Drezner // Annals of Operations Research.- 122 (1-4).- 2003.- P. 21–42, doi 10.1023/A:1026130003508

89. Antamoshkin, A.N. Random Search Algorithm for the p-Median Problem / A.N. Antamoshkin, L.A. Kazakovtsev // Informatica.-2013.-Vol. 37, issue 3.- P.267–278.
90. Ausiello, G. Local Search, Reducibility and Approximability of NP-optimization Problems / G. Ausiello, M. Protasi // Information Processing Letters.-1995.- Vol.54.- P. 73-79.
91. Baldi, P. DNA Microarrays and Gene Expression /P. Baldi., G. Hatfield.-[s.l.]: Cambridge University Press.-2002.- P.208.
92. Bandyopadhyay, S. An evolutionary technique based on K-Means algorithm for optimal clustering / S. Bandyopadhyay, U. Maulik // Information Sciences.-2002.-Vol. 146.- P.221-237.
93. Bartlett M. S., M a c d o n a l d P. D. M., «Least-squares» estimation of distribution mixtures. Nature (Engl.), 1968, 217, № 5124, 195—196.
94. Behboodian, J. On a mixture of normal distributions. Biometrika, 1970, 57, № 1, 215—217.
95. Behboodian, J. On the distribution of a symmetric statistic from a mixed population. Technometrics. 1972, 14, № 4, 919-923.
96. Bellman, R.E. Dynamic Programming / R.E.Bellman.- NJ,Princeton: Princeton University Press.- 1957.- P.392.
97. Bhatia, S. Conceptual clustering in information retrieval / S. Bhatia, J. Deogun // IEEE Trans. Systems Man Cybernet.- 1998.- Vol. 28 (B).- P. 427–436.
98. Blei, D.M.Latent dirichlet allocation / D.M. Blei, A.Y. Ng, M.I. Jordan // J. Machine Learn. Res.- 2003.- Vol.3.- P. 993–1022.
99. Blischke, W. R. Estimating the parameters of mixtures of binomial distributions. J. Amer. Statist. Assoc, 1964, 59, № 306, 510—528.
100. Boes, D. C. Minimax unbiased estimator of mixing distribution for finite mixtures. Sankhya. Indian J. Statist., 1967, A29, № 4, 417—420.

101. Boese, K.D. A new adaptive multi-start technique for combinatorial global optimizations / K.D. Boese, A.B. Kahng, S. Muddu // Oper. Res. Lett.- 1994.- Vol. 16, No 2.- P. 101-114.
102. Bozkaya, B.A. Genetic Algorithm for the p-Median Problem / B. Bozkaya, J. Zhang, E.Erkut // Facility Location: Applications and Theory / Z. Drezner, H. Hamacher [eds.].-New York: Springer.-2002.- P. 179-205.
103. Broniatowski M., Celeux G., Diebolt J. Reconnaissance de m´elanges de densit´es par un algorithme d'apprentisage probabiliste. – in: E. Diday, M. Jambu, L. Lebart, J.-P. Pag`es and R. Tomasone (Eds.) Data Analysis and Informatics, III, North Holland, Amsterdam, 1983, p. 359-373.
104. Calinski T.,Harabasz J. A dendrite method for cluster analysis // Communications in Statistics, 1974. Vol. 3. P. 1-27.doi:10.1080/03610927408827101
105. Celeux G, Govaert A. Classification EM Algorithm for Clustering and Two Stochastic Versions. Rapport de Recherche de l'INRIA RR-1364. Centrede Rocquencourt. 1991
106. Celeux G., Diebolt J. Reconnaissance de m´elanges de densit´e et classification. Un algorithme d'apprentisage probabiliste: l'algorithme SEM. Rapport de Recherche de l'INRIA RR-0349. Centre de Rocquencourt. 1984.
107. Celeux G., Diebolt J. The SEM algorithm: a probabilistic teacher algorithm derived from the EM algorithm for the mixture problem. – Computational Statistics Quarterly, 1985, vol. 2, No. 1, p. 73-82.
108. Celeux G., Govaert G. A classification EM algorithm for clustering and two stochastic versions. – Computational Statistics and Data Analysis, 1992, vol. 14, p. 315-332.
109. Chatteriee S. K., Rank approach to the multivariate two-population mixture problem. J. Multivar. Anal., 1972, 2, № 3, 261—281.
110. Chiou, Y. Genetic clustering algorithms / Y. Chiou, L.W. Lan // European Journal of Operational Research.- 2001.- Vol. 135.- P. 413-427.

111. Choi K., B u l e r e n W- G., An estimation procedure for mixtures of distributio'ns. J. Roy. Statist- Soc, '1968, B30, № 3, 444-460.

112. Choi K., Estimators for the parameters of a finite mixture of distributions. Ann. Inst. Statist. Math., 1969, 21, № 1, 107-116 (РЖМат, 1970, 2В155)

113. Connell, S.D. Writer adaptation for online handwriting recognition / S.D. Connell, A.K. Jain//IEEE Trans. Pattern Anal. Machine Intell.- 2002.- Vol.24, issue 3, P.329–346.

114. Davies, D.L., Bouldin, D.W. A Cluster Separation Measure // IEEE Transactions on Pattern Analysis and Machine Intelligence, 1979. PAMI-1 (2).P.224–227.

115. Day N. E., Estimating the components of mixture of normal distributions. Biometrika, 1969, 56, № 3, 463—474.

116. Deely J. J., K r u s e R- L., Construction of sequences estimating the mixing distribution. Ann. Math. Stat., 1968, 39, № 1, 286—288.

117. Dempster A., Laird N., Rubin D. Maximum likelihood estimation from incomplete data. – Journal of the Royal Statistical Society, Series B, 1977, vol. 39, p. 1-38.

118. Dick N. P., B o w d e n D. C, Maximum likelvhood estimation for mixtures of two normal distributions. Biometrics, 1973, 29, № 4, 781—790

119. Drezner, Z. The fortified Weiszfeld algorithm for solving the Weber problem / Z. Drezner // IMA Journal of Management Mathematics.-2013.-Vol.26.- P.1-9. DOI: 10.1093/imaman/dpt019

120. Duda, R. Pattern Classification, second ed. / R. Duda., P. Hart, D. Stork.- New York:John Wiley and Sons.- 2001.- P.680

121. Engelman L., i l - h a r t i g a n J. A., Percentage points of a test for clusters. J. Amer. Statist. Assoc, 1969, 64, № 323, 1647—1648.

122. Eremeev, A.V. Optimal recombination in genetic algorithms for combinatorial optimization problems, part 1 / A.V. Eremeev, J.V. Kovalenko // Yugoslav

Journal of Operations Research.- 2014.- Vol.24, issue 1.- P. 1-20, DOI:10.2298/YJOR130731040E.

123. Ester, M. Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise / M. Ester, H.-P. Kriegel, J. Sander, X. Xu // KDD-96 Proceedings.-[s.l.]:[s.n.].- 1996.- P. 226-231.
124. Frank, I.E. Data Analysis Handbook / I.E. Frank, R. Todeschini.-[s.l.]: Elsevier Science Inc.- 1994.- P. 227–228.
125. Friedman H. P., R u b i n J., On some invariant criteria for grouping data. J. Amer. Statist. Assoc, 1967, 62, № 320, 1159—1178.
126. Fryer J. G-, R o b e r t s o n C A., A comparison of some methods for estimating mixed normal distributions. Biometrika, 1972, 59, № 3, 639—648.
127. Fukunaga K., K o o n t z W. L. G., A criterion and an algorithm for grouping data. IEEE Trans. Comput., 1970, 19, № 10, 917—923.
128. Ganti, V. Clustering large datasets in arbitrary metric spaces / V. Ganti, R. Ramakrishnan, J. Gehrke, A. Powell, J. French. // Proc. 15th Int. Conf. Data Engineering.- [s.l.]:[s.n.].- 1999.- P. 502-511.
129. Goldberg, D.E. Genetic algorithm in search, optimization and machine learning / D.E. Goldberg.-MA: Addison-Wesley.-1989.-P. 432.
130. Gridgeman N. T., A comparison of two methods of analysis of mixtures of normal distributions- Technometrics, 1970, 12, № 4, 823—833.
131. Hansen P. Solving large p-median clustering problems by primaldual variable neighborhood search / P. Hansen, J. Brimberg, D. Urosevic, N. Mladenovic // Data Mining and Knowledge Discovery.- 2009.- 19, No. 3.- P. 351–375.
132. Hansen, Lars Peter (2002). "Method of Moments". In Smelser, N. J.; Bates, P. B. International Encyclopedia of the Social and Behavior Sciences. Oxford: Pergamon.
133. Hansen, P. Variable Neighborhood Search / P. Hansen, N. Mladenovic // Search Methodology /E.K.Bruke, G.Kendall [eds.].-Springer US.- 2005.- P. 211-238, doi: 10.1007/0-387-28356-0_8.

134. Hansen, P. Variable neighborhood search: principles and applications / P. Hansen, N. Mladenovic // Eur. J. Oper. Res.- 2001.- Vol.130.- P.449–467.
135. Hartigan, J.A. Clustering Algorithms // New York: Wiley. 1975. 369 P.
136. Hasmer, D.W. On MLE the parameters of a mixture ol two normal distributions when the sample size lis small- Communic. Stat., 1973, 1, № 3, 217—225
137. Hasselblad, V. Estimation of finite mixtures of distributions from the exponential family, J. Amer. Statist- Assoc, 1969, 64, № 328, 1459—1471.
138. Hasselblad, V. Estimation of parameters for a mixture of normal distributions. Technometrics, 1966, 8, № 3, 431—446.
139. Hawkins, R. H. A note on multiple solutions to the mixed distribution problem. Technometrics, 1972, 14, № 4, 973—976.
140. Holland, J. Adaptation in Natural and Artificial Systems. Ann / J. Holland.- Arbor: University of Michigan Press.- 1975.- P.183.
141. Hosmer, D.W., Jr. A comparison of iterative maximum likelihood estimates of the parameters of a mixture of two normal distributions under three different types of sample. Biometrics, 1973, 29, № 4, 761— 770.
142. Hu, J. Statistical methods for automated generation of service engagement staffing plans / J. Hu, B.K. Ray, M. Singh // IBM J. Res. Dev. 2007. Vol. 51, issue 3. P. 281–293.
143. Iwayama, M.Cluster-based text categorization: A comparison of category search strategies / M. Iwayama, T. Tokunaga // Proc. 18th ACM Internat. Conf. on Research and Development in Information Retrieval.- 1995.- P. 273–281.
144. Jain, A.K. Data clustering: 50 years beyond K-means /A.K. Jain // Pattern Recognition Letters.-2010.-Vol.31.- P. 651-666.
145. Jain, A.K.Image segmentation using clustering /A.K. Jain, P. Flynn // Advances in Image Understanding.-IEEE Computer Society Press.-1996.- P.65–83.
146. John E., Bnyesian estimation of mixture distributions. Ann. Math. Stat, 1968, 39, № 4, 1289—1302

147. John S., On identifying the population of origin of each observation in a mixture of observations from two gamma populations. Technometrics, 1970, 12, M> 3, 565—568.

148. John S., On identifying the population of origin of each observation in a mixture of observations from two normal populations. Technometrics, 1970, 12, № 3, 553—563.

149. Kabir A. iB. M., Estimation of parameters of a finite mixture of distributions. J. Roy. Statist. Soc, 1968, B30, № 3, 472—482.

150. Kaski E-, K r y s i c k i W-, Die Parameter schatzung einer Mischung von zwei JLaplacischen Verteilungen (in lallgemetaem Fall). Prace Mat., 1967, 11, ser. 1, 23—31

151. Kazakovtsev L.A. Fast Deterministic Algorithm for EEE Components Classification / L.A.Kazakovtsev, A.N.Antamoshkin, I.S.Masich // IOP Conf. Series: Materials Science and Engineering.–2015.–Vol.94.–article ID 012015, 10 P. DOI: 10.1088/1757-899X/04/1012015

152. Kazakovtsev L.A. Modified Genetic Algorithm with Greedy Heuristic for Continuous and Discrete p-Median Problems / L.A. Kazakovtsev, V.I. Orlov, A.A. Stupina, V.L. Kazakovtsev //FactaUniversitatis (Nis) Series Mathematics and Informatics.- 2015.- Vol. 30, No. 1.- P. 89-106.

153. Kazakovtsev L.A., Antamoshkin A.N., Fedosov V.V. Greedy heuristic algorithm for solving series of EEE components classification problems // IOP Conference Series: Materials Science and Engineering. 2016. Vol. 122.ArticleID 012011. 7 P. DOI: 10.1088/1757-899X/122/1/012011.

154. Kazakovtsev, L. 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.

155. Kirkpatrick, S. Optimization by simulated annealing / S. Kirkpatrick, C.D. Gelatt, M.P. Vecchi // Science.-1983.-Vol. 220(4598).- P. 671–680.

156. Kohonen, T. Self-Organization and Associative Memory, 3rd ed. / T. Kohonen // Springer information sciences series.-New York:Springer-Verlag.- 1989.- P.312.
157. Krishna, K. Genetic K-means algorithm / K. Krishna, M. Murty // IEEE Transaction on System, Man and Cybernetics - Part B.- 1999.- Vol.29.- P. 433-439.
158. Krysicki W., Estimation of the parameters of the mixture of an arbitrary number of exponential distributions. Demonstr. math., 1972, 4, № 3, 175—183.
159. Krzanowski W., Lai Y. A criterion for determining the numberofgroups in a dataset using sum of squares clustering // Biometrics, 1985. No. 44.P. 23–34.
160. Lange, T. Learning with constrained and unlabelled data /T. Lange, M.H. Law, A.K. Jain, J. Buhmann // IEEE Comput. Soc. Conf. Comput. Vision Pattern Recognition.- 2005.- Vol.1.- P.730-737.
161. Levanova, T.V. Algorithms of Ant System and simulated annealing for the p-median problem / T.V. Levanova, M.A. Loresh // Automation and remote control. -2004. -Vol.65, issue 30. -P. 431-438.
162. Li, W. Pachinko allocation: Dag-structured mixture models of topic correlations / W. Li, A. McCallum // Proc. 23rd Internat. Conf. on Machine Learning.- [s.l.]:[s.n.].- 2006.- P. 577–584.
163. Likas, A. The global k-means clustering algorithm / A. Likas, M. Vlassis, J. Verbeek // Pattern Recognition.-2003.-Vol.36.- P. 451-461.
164. Lu, Y. Incremental genetic k-means algorithm and its application in gene expression data analysis / Y. Lu, S. Li, F. Fotouhi, Y. Deng, S. Brown // BMC Bioinformatics.- 2004. - Vol.5, No.1.- Article 172.- [Электронный ресурс] режим доступа DOI: 10.1186/1471-2105-5-172 (дата обращения 15.04.2017).
165. Mac Queen J., Some methods for classification and analysis of multivariate observations. Proc. 5th Berkeley Sympos. Math. Statist, and Probabil, 1965—1966. Vol- 1- Berkeley—Los Angeles, 1967, 281—297.
166. Maulik, U. Genetic Algorithm-Based Clustering Technique / U. Maulik, S. Bandyopadhyay // Pattern Recognition.-2000.-Vol. 33.- P. 1455-1465.

167. Mayer L. S., A method of cluster analysis when there exist multiple indicators of a theoretic concept Biometrics, 1971, 27, № 1, 143—155
168. McLachlan, G.L. Mixture Models: Inference and Applications to Clustering / G.L. McLachlan, K.E. Basford.-New York: Marcel Dekker.-1987.- P.253.
169. Medgyessy P., V a r g a L., Gauss-fiiggveny keverekek numerikus felbontasara szolgalo egyik eljaras javitasarol. I—4. Magy. tud. akad. Mat. 6s fiz. tud. oszt. kozl., 196i3, 18, № 1, 31—39.
170. Meeden G., iBayes estimation of the mixing distribution, the discrete case, Ann. Math. Stat., 1972, 43, № 6, 1993—1999.
171. Mladenovic, N. Variable neighborhood search / N. Mladenovic, P. Hansen // Comput. Oper. Res. 1997.- Vol.24.- P.1097–1100.
172. Mohd, W.M.B.W. An Improved Parameter less Data Clustering Technique based on Maximum Distance of Data and Lioyd k-means Algorithm / W.M.B.W. Mohd, A.H. Beg, T. Herawan, K.F. Rabbi // First World Conference on Innovation and Computer Sciences (INSODE 2011).- [s.l.]:[s.n.].-2012.- Vol.1.- P.7–371, DOI: 10.1016/j.protcy.2012.02.076
173. Muhlenbein, H. The Equation for Response to Selection and Its Use for Prediction / H. Muhlenbein // Evolutionary Computation.-1998.-Vol.5, No. 3.- P. 303-346.
174. Neema, M.N. New Genetic Algorithms Based Approaches to Continuous p-Median Problem / M.N. Neema, K.M. Maniruzzaman, A. Ohgai // Netw. Spat. Econ.- 2011.- Vol.11.- P.83–99, DOI:10.1007/s11067-008-9084-5.
175. Orlov V.I, Stashkov D.V., Kazakovtsev L.A., Stupina A.A. Fuzzy clustering of EEE components for space industry // IOP Conference Series: Materials Science and Engineering. 2016. Vol. 155. P. 012026. doi: 10.1088/1757-899X/155/1/012026.
176. Parzen, E. On Estimation of a Probability Density, Function and Mode / E. Parzen // II IEEE Transactions on Information Theory.- 1982.- Vol. 4, No. 6.- P. 663-666.

177. Parzen, E. On the Estimation of Probability Density Function and the Mode / E. Parzen / E. Parzen // Ann. Math. Statist.- 1962.- Vol. 33.- P. 1065.
178. Patrick E. A., C o s t e l l o J. P., On unsupervised estimation algorithms. IEEE Trans. Inform- Theory, 1970, 16, № 5, 556-569.
179. Preston P. F., Estimating the mixing distribution by piece-wise polynomialarcs. Austnal. J. Statist., 1971, 13, № 2, 64—76.
180. Rand W. M., Objective oriteria for the evaluation of oliusterinigmethods. J. Amer. Statist. Assoc, 1971, 66, N» 336, 846—850.
181. Reeves, C.R. Genetic algorithms for the operations researcher / C.R. Reeves // INFORMS Journal of Computing.-1997.-Vol.9, issue 3.- P.231-250.
182. Roberts, S.J. Minimum-entropy data clustering using reversible jump Markov chain Monte Carlo / S.J. Roberts, C. Holmes, D. Denison // Proc. Internat. Conf. Artificial Neural Networks.- [s.l]:[s.n.].- 2001.- P. 103–110.
183. Rousseeuw, P. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis / P. Rousseeuw // Journal of Computational and Applied Mathematics.- 1987.- Vol. 20.- P. 53-65.
184. Sage, A.P. System Identification / A.P. Sage, J.L. Melsa.-[s.l.]:Chu I.- 1971.- P.221.
185. Schnell P., Eine Methode zur Auffindung von gruppen. Biometr. Z,, 1964,. 6, № 1, 47—48
186. Schwarz, G. Estimating the Dimension of a Model // Annals of Statistics. 6 (1978), no. 2, 461-464. doi:10.1214/aos/1176344136.
187. Scott A. J,, Symons M. J., Clustering methods based on likelihood ratio criteria. Biometrics, 1971, 27, № 2, 387—397.
188. Semi-Supervised Learning / O. Chapelle, B. Schoelkopf.,A.Zien (Eds.).- Cambridge:MIT Press.- 2006.- P.508
189. Sheng, W. A genetic k-medoids clustering algorithm / W. Sheng, X. Liu //Journal of Heuristics.-2006.-Vol.12, No.6.- P. 447-466.

190. Shi, J. Normalized cuts and image segmentation / J. Shi, J. Malik // IEEE Trans. Pattern Anal. Machine Intell.- 2000.- Vol.22.- P. 888–905.
191. Singh M. P., A note on generalized inflated binomial distribution. Sankhya. Indian J. Statist., 1966, A28, № 1, 99.
192. Sleeba J., On analyzing mixed samples. J. Amer. Statist. Assoc, i970, 65, № 330, 755-762.
193. Slonim, N. Document clustering using word clusters via the information bottleneck method / N. Slonim, N. Tishby // ACM SIGIR 2000.- 2000.- P. 208–215.
194. Stanat D. F., Nonsupervised pattern recognition through the decomposition of probability functions. Doct. diss. Univ. Mich,, 1966, 60 pp. Dissert. Abstrs., 1967, B27, № 7, 2452.
195. Still, S. Geometric Clustering using the Information Bottleneck method / S. Still, W. Bialek, L. Bottou // Advances In Neural Information Processing Systems 16 / Eds.:S. Thrun,L. Saul, and B. Scholkopf.- Cambridge:MIT Press.- 2004 [Электронный ресурс] Режим доступа URL http://papers.nips.cc/paper/2361-geometric-clustering-using-the-information-bottleneck-method.pdf (дата обращения 16.09.2016)
196. Tabachnick, B.G. Using Multivariate Statistics, fifth ed. / B.G.Tabachnick, L.S. Fidell.- Boston:Allyn and Bacon.- 2007.- P.980
197. Thionet P., Note sur les melanges de certaines distributions de probabilit y .iPubls. Inst, statist. Univ. Paris, 1966, 15, № 1, 61—80.
198. Tibshirani, R., Walther, G, Hastie, T. Estimating the number of clusters in a data set via the gap statistic // Journal of the Royal Statistical Society, 2001.Vol. 63.P. 411–423.
199. Tishby, N. The information bottleneck method / N. Tishby, F.C. Pereira, W. Bialek // Proc. 37th Allerton Conf. on Communication, Control and Computing.- Monticello:[s.n.].-1999.- P. 368–377.

200. Tukey, J.W. Exploratory Data Analysis / J.W. Tukey.- Addison-Wesley.- 1977.- P.688

201. Urbakh V. Yu., A discriminant method of clustering. J. Multivar. Anal., 1972, 2, № 3, 249—260.

202. Vidyasagar, M. Statistical learning theory and randomized algorithms for control / M. Vidyasagar // IEEE Control Systems. -1998.-No. 12. -P. 69-85.

203. Weber, A. Ueber den Standort der Industrien, Erster Teil: Reine Theorie des Standortes/ A. Weber, [Verlag von] J. C. B. Mohr.- Tuebingen: Mohr.- 1922.- 209 P.

204. Weiszfeld, E. Sur le point sur lequel la somme des distances de n points donnes est minimum/ E. Weiszfeld // Tohoku Mathematical Journal.-1937.-Vol. 43, No.1.- P.335–386.

205. Welling, M. Exponential family harmoniums with an application to information retrieval / M. Welling, M. Rosen-Zvi, G. Hinton // Adv. Neural Inform. Process. Systems.- 2005.- Vol.17.- P. 1481–1488.

206. Wolfe J. H,, Pattern clustering by multivariate mixture analysis. Multivariate Behavior. Res., 1970, 5, 329

207. Yakowitz S. J., A consistent estimator for the identification of finite mixtures. Ann, Math. Stat., 1969, 40, № 5, 1728—1735.

208. Young T. Y., C oral up pi G., Stochastic estimation of a mixture of normal density functions using an information criterion. IEEE Trans. Inform. Theory, 1970, 16, № 3, 258—263.

ПРИЛОЖЕНИЕ А. Сравнительный анализ вычислительных экспериментов различных алгоритмов

В таблицах А.1, ..., А.20 приведены результаты вычислительных экспериментов для задач разделения смеси распределений, выполненных для новых генетических алгоритмов (GA-ONE, GA-FULL, GA-RAND), новых алгоритмов поиска в чередующихся окрестностях (VNS-1, VNS-2, VNS-3) и известных алгоритмов (EM-алгоритма и его модификаций: CEM (классификационный), SEM (стохастический)).

Выполнено сравнение результатов работы новыхи известных алгоритмов по значению целевой функции. Статистическая значимость результатов проверена гипотезой о неравенстве математических ожиданий результатов новых и известных алгоритмов H:$f_{\text{лучший из новых алгоритмов}} > f_{\text{лучший из известных модификаций EM}}$.

Обозначим через $\overline{m}_1$ и $\overline{m}_2$ вычисленные оценки математических ожиданий значений целевой функции, полученных сравниваемымиалгоритмами при nзапусках;$x_1, \ldots, x_n$и$y_1, \ldots, y_n$–значения целевой функции, полученные при каждом запуске. Для проверки гипотезы $H : \overline{m}_1 \neq \overline{m}_2$вводим t-статистику (критерий Стьюдента):

$$z = \frac{\overline{m}_1 - \overline{m}_2}{\overline{\sigma}\sqrt{1/n_1 + 1/n_2}}, \text{ где } \overline{\sigma}^2 = \frac{1}{n_1 + n_2 - 2} \cdot \left(\sum_{i=1}^{n} \left(x_i - \overline{m}_1\right)^2 + \sum_{i=1}^{n} \left(y_i - \overline{m}_1\right)^2 \right).$$

Для проверки гипотезы неравенства математических ожиданий проверяем значения неравенства $|z| < t_{\alpha=1\%,\ 2n-2}$, где $t_{\alpha=1\%,\ 2n-2}$ - пороговое значение статистики Стьюдента. Результаты из таблиц показывают, что при уровне значимости 1% различия математических ожиданий статистически значимы.

Таблица А.1 – сравнительные результаты работы ГА с жадной эвристикой для задач кластеризации и размещения.

Характеристика набора данных, задачи и алгоритма			Значение
Наименование набора данных			Chess (UCI), булевы
Число векторов данных			N=3197
Размерность пространства			d=50
Время работы алгоритмов			t=10 мин.
Число запусков алгоритмов			30
Число кластеров			k=30
Алгоритм		Показатель	Значение
Модификации EM	EM	среднее	-30 525,04*
		σ	0,00
	CEM	среднее	-30 564,13
		σ	32,29
	SEM	среднее	-30 560,70
		σ	46,39
	Лучший рез-т модификаций	среднее	-30 525,04**
		σ	0,00
GA-ONE		среднее	-30 532,44
		σ	19,55
(в сравнении GA-ONE с лучшим из EM)		среднее	-0,0243%
		σ	+0,0000%
Гипотеза о неравенстве мат. ожиданий $H: f^*_{GA\text{-}ONE} > f^*_{лучший\ EM}$		Статистика Стьюдента	-2,94
		Порог статистики при α=1%	2,66
		Принятие гипотезы	нет, преимущества нет
GA-FULL		среднее	-30 581,09
		σ	1,83
(в сравнении GA-FULL с лучшим из EM)		среднее	-0,1836%
		σ	+0,0000%
Гипотеза о неравенстве мат. ожиданий $H: f^*_{GA\text{-}FULL} > f^*_{лучший\ EM}$		Статистика Стьюдента	-236,68
		Порог статистики при α=1%	2,66
		Принятие гипотезы	нет, преимущества нет
GA-RAND		среднее	-30 554,67
		σ	28,69
(в сравнении GA-RAND с лучшим из EM)		среднее	-0,0971%
		σ	+0,0000%
Гипотеза о неравенстве мат. ожиданий $H: f^*_{GA\text{-}RAND} > f^*_{лучший\ EM}$		Статистика Стьюдента	-7,99
		Порог статистики при α=1%	2,66
		Принятие гипотезы	нет, преимущества нет

Примечание: * -лучший результат семейства алгоритмов (EM, GA, VNS), **- абсолютно лучший результат среди всех алгоритмов.

Таблица А.2 – сравнительные результаты работы VNS с жадной эвристикой для задач кластеризации и размещения.

Характеристика набора данных, задачи и алгоритма			Значение
Наименование набора данных			Chess (UCI), булевы
Число векторов данных			N=3197
Размерность пространства			d=50
Время работы алгоритмов			t=10 мин.
Число запусков алгоритмов			30
Число кластеров			k=30
Алгоритм		Показатель	Значение
Модификации EM	EM	среднее	-30 525,04*
		σ	0,00
	CEM	среднее	-30 564,13
		σ	32,29
	SEM	среднее	-30 560,70
		σ	46,39
	Лучший рез-т модификаций	среднее	-30 525,04**
		σ	0,00
VNS-1		среднее	-30 554,67
		σ	28,69
(в сравнении VNS-1 с лучшим из EM)		среднее	-0,0971%
		σ	+0,0000%
Гипотеза о неравенстве мат. ожиданий $H:f^*_{VNS\text{-}1}>f^*_{лучший\ EM}$		Статистика Стьюдента	-8,00
		Порог статистики при α=1%	2,66
		Принятие гипотезы	нет, преимущества нет
VNS-2		среднее	-30 539,85
		σ	25,43
(в сравнении VNS-2 с лучшим из EM)		среднее	-0,0485%
		σ	+0,0000%
Гипотеза о неравенстве мат. ожиданий $H:f^*_{VNS\text{-}2}>f^*_{лучший\ EM}$		Статистика Стьюдента	-4,51
		Порог статистики при α=1%	2,66
		Принятие гипотезы	нет, преимущества нет
VNS-3		среднее	-30 580,60
		σ	0,00
(в сравнении VNS-3 с лучшим из EM)		среднее	-0,1820%
		σ	+0,0000%
Гипотеза о неравенстве мат. ожиданий $H:f^*_{VNS\text{-}3}>f^*_{лучший\ EM}$		Статистика Стьюдента	-15 527 490,30
		Порог статистики при α=1%	2,66
		Принятие гипотезы	нет, преимущества нет

Примечание: * -лучший результат семейства алгоритмов (EM, GA, VNS), **- абсолютно лучший результат среди всех алгоритмов.

Таблица А.3 – сравнительные результаты работы ГА с жадной эвристикой для задач кластеризации и размещения.

Характеристика набора данных, задачи и алгоритма			Значение
Наименование набора данных			Ionosphere (UCI), норм.
Число векторов данных			N= 351
Размерность пространства			d= 35
Время работы алгоритмов			t= 2 мин.
Число запусков алгоритмов			30
Число кластеров			k= 10
Алгоритм		Показатель	Значение
Модификации EM	EM	среднее	-871,41*
		σ	15,79
	CEM	среднее	-893,44
		σ	7,88
	SEM	среднее	-879,95
		σ	15,19
	Лучший рез-т модификаций	среднее	-871,41
		σ	15,79
GA-ONE		среднее	-824,85**
		σ	4,47
(в сравнении GA-ONE с лучшим из EM)		среднее	+5,3430%
		σ	-71,6625%
Гипотеза о неравенстве мат. ожиданий $H: f^*_{GA\text{-}ONE} > f^*_{лучший\ EM}$		Статистика Стьюдента	21,98
		Порог статистики при α=1%	2,66
		Принятие гипотезы	да, преимущество статистически значимо
GA-FULL		среднее	-960,32
		σ	23,50
(в сравнении GA-FULL с лучшим из EM)		среднее	-10,2034%
		σ	+48,8666%
Гипотеза о неравенстве мат. ожиданий $H: f^*_{GA\text{-}FULL} > f^*_{лучший\ EM}$		Статистика Стьюдента	-24,33
		Порог статистики при α=1%	2,66
		Принятие гипотезы	нет, преимущества нет
GA-RAND		среднее	-832,83
		σ	14,80
(в сравнении GA-RAND с лучшим из EM)		среднее	+4,4270%
		σ	-6,2572%
Гипотеза о неравенстве мат. ожиданий $H: f^*_{GA\text{-}RAND} > f^*_{лучший\ EM}$		Статистика Стьюдента	13,81
		Порог статистики при α=1%	2,66
		Принятие гипотезы	да, преимущество статистически значимо

Примечание: * -лучший результат семейства алгоритмов (EM, GA, VNS), **- абсолютно лучший результат среди всех алгоритмов.

Таблица А.4 – сравнительные результаты работы VNS с жадной эвристикой для задач кластеризации и размещения.

Характеристика набора данных, задачи и алгоритма			Значение
Наименование набора данных			Ionosphere (UCI), норм.
Число векторов данных			N= 351
Размерность пространства			d= 35
Время работы алгоритмов			t= 2 мин.
Число запусков алгоритмов			30
Число кластеров			k= 10
Алгоритм		Показатель	Значение
Модификации EM	EM	среднее	-871,41*
		σ	15,79
	CEM	среднее	-893,44
		σ	7,88
	SEM	среднее	-879,95
		σ	15,19
	Лучший рез-т модификаций	среднее	-871,41
		σ	15,79
VNS-1		среднее	-847,46*
		σ	23,97
(в сравнении VNS-1 с лучшим из EM)		среднее	+2,7475%
		σ	+51,8589%
Гипотеза о неравенстве мат. ожиданий $H: f^*_{VNS\text{-}1} > f^*_{лучший\ EM}$		Статистика Стьюдента	6,46
		Порог статистики при α=1%	2,66
		Принятие гипотезы	да, преимущество статистически значимо
VNS-2		среднее	-849,35
		σ	28,07
(в сравнении VNS-2 с лучшим из EM)		среднее	+2,5307%
		σ	+77,8391%
Гипотеза о неравенстве мат. ожиданий $H: f^*_{VNS\text{-}2} > f^*_{лучший\ EM}$		Статистика Стьюдента	5,30
		Порог статистики при α=1%	2,66
		Принятие гипотезы	да, преимущество статистически значимо
VNS-3		среднее	-849,35
		σ	28,07
(в сравнении VNS-3 с лучшим из EM)		среднее	+2,5307%
		σ	+77,8391%
Гипотеза о неравенстве мат. ожиданий $H: f^*_{VNS\text{-}3} > f^*_{лучший\ EM}$		Статистика Стьюдента	5,30
		Порог статистики при α=1%	2,66
		Принятие гипотезы	да, преимущество статистически значимо

Примечание: * -лучший результат семейства алгоритмов (EM, GA, VNS), **- абсолютно лучший результат среди всех алгоритмов.

Таблица А.5 – сравнительные результаты работы ГА с жадной эвристикой для задач кластеризации и размещения.

Характеристика набора данных, задачи и алгоритма			Значение
Наименование набора данных			Mopsi (UCI)
Число векторов данных			N= 6014
Размерность пространства			d= 2
Время работы алгоритмов			t= 40 мин.
Число запусков алгоритмов			30
Число кластеров			k= 20
Алгоритм		Показатель	Значение
Модификации EM	EM	среднее	39 268,66
		σ	9 967,64
	CEM	среднее	48 424,17*
		σ	237,30
	SEM	среднее	36 272,22
		σ	9 619,60
	Лучший рез-т модификаций	среднее	48 424,17
		σ	237,30
GA-ONE		среднее	49 288,89
		σ	390,47
(в сравнении GA-ONE с лучшим из EM)		среднее	+1,7857%
		σ	+64,5512%
Гипотеза о неравенстве мат. ожиданий $H:f^*_{GA\text{-}ONE}>f^*_{лучший\ EM}$		Статистика Стьюдента	14,66
		Порог статистики при α=1%	2,66
		Принятие гипотезы	да, преимущество статистически значимо
GA-FULL		среднее	50 443,36
		σ	115,25
(в сравнении GA-FULL с лучшим из EM)		среднее	+4,1698%
		σ	-51,4301%
Гипотеза о неравенстве мат. ожиданий $H:f^*_{GA\text{-}FULL}>f^*_{лучший\ EM}$		Статистика Стьюдента	59,29
		Порог статистики при α=1%	2,66
		Принятие гипотезы	да, преимущество статистически значимо
GA-RAND		среднее	50 499,90**
		σ	60,91
(в сравнении GA-RAND с лучшим из EM)		среднее	+4,2866%
		σ	-74,3305%
Гипотеза о неравенстве мат. ожиданий $H:f^*_{GA\text{-}RAND}>f^*_{лучший\ EM}$		Статистика Стьюдента	65,63
		Порог статистики при α=1%	2,66
		Принятие гипотезы	да, преимущество статистически значимо

Примечание: * -лучший результат семейства алгоритмов (EM, GA, VNS), **- абсолютно лучший результат среди всех алгоритмов.

Таблица А.6 – сравнительные результаты работы VNS с жадной эвристикой для задач кластеризации и размещения.

Характеристика набора данных, задачи и алгоритма			Значение
Наименование набора данных			Mopsi (UCI)
Число векторов данных			N= 6014
Размерность пространства			d= 2
Время работы алгоритмов			t= 40 мин.
Число запусков алгоритмов			30
Число кластеров			k= 20
Алгоритм		Показатель	Значение
Модификации EM	EM	среднее	39 268,66
		σ	9 967,64
	CEM	среднее	48 424,17*
		σ	237,30
	SEM	среднее	36 272,22
		σ	9 619,60
	Лучший рез-т модификаций	среднее	48 424,17
		σ	237,30
VNS-1		среднее	50 291,12
		σ	122,86
(в сравнении VNS-1 с лучшим из EM)		среднее	+3,8554%
		σ	-48,2251%
Гипотеза о неравенстве мат. ожиданий $H:f^*_{VNS\text{-}1}>f^*_{лучший\ EM}$		Статистика Стьюдента	54,12
		Порог статистики при α=1%	2,66
		Принятие гипотезы	да, преимущество статистически значимо
VNS-2		среднее	50 311,64
		σ	104,68
(в сравнении VNS-2 с лучшим из EM)		среднее	+3,8978%
		σ	-55,8857%
Гипотеза о неравенстве мат. ожиданий $H:f^*_{VNS\text{-}2}>f^*_{лучший\ EM}$		Статистика Стьюдента	56,37
		Порог статистики при α=1%	2,66
		Принятие гипотезы	да, преимущество статистически значимо
VNS-3		среднее	50 370,82*
		σ	50,99
(в сравнении VNS-3 с лучшим из EM)		среднее	+4,0200%
		σ	-78,5135%
Гипотеза о неравенстве мат. ожиданий $H:f^*_{VNS\text{-}3}>f^*_{лучший\ EM}$		Статистика Стьюдента	62,13
		Порог статистики при α=1%	2,66
		Принятие гипотезы	да, преимущество статистически значимо

Примечание: * -лучший результат семейства алгоритмов (EM, GA, VNS), **- абсолютно лучший результат среди всех алгоритмов.

Таблица А.7 – сравнительные результаты работы ГА с жадной эвристикой для задач кластеризации и размещения.

Характеристика набора данных, задачи и алгоритма			Значение
Наименование набора данных			Europe (UCI), некорр.
Число векторов данных			N= 169308
Размерность пространства			d= 2
Время работы алгоритмов			t= 240 мин.
Число запусков алгоритмов			30
Число кластеров			k= 30
Алгоритм		Показатель	Значение
Модификации EM	EM	среднее	-3 653 148,61*
		σ	2 634,66
	CEM	среднее	-
		σ	-
	SEM	среднее	-3 654 493,58
		σ	4 348,55
	Лучший рез-т модификаций	среднее	-3 653 148,61
		σ	2 634,66
GA-ONE		среднее	-3 651 916,99
		σ	3 465,70
(в сравнении GA-ONE с лучшим из EM)		среднее	+0,0337%
		σ	+31,5424%
Гипотеза о неравенстве мат. ожиданий $H: f^*_{GA\text{-}ONE} > f^*_{лучший\ EM}$		Статистика Стьюдента	2,19
		Порог статистики при α=1%	2,66
		Принятие гипотезы	нет, преимущество статистически не значимо
GA-FULL		среднее	-3 643 277,02*
		σ	2 122,94
(в сравнении GA-FULL с лучшим из EM)		среднее	+0,2702%
		σ	-19,4227%
Гипотеза о неравенстве мат. ожиданий $H: f^*_{GA\text{-}FULL} > f^*_{лучший\ EM}$		Статистика Стьюдента	22,60
		Порог статистики при α=1%	2,66
		Принятие гипотезы	да, преимущество статистически значимо
GA-RAND		среднее	-3 648 896,16
		σ	5 151,49
(в сравнении GA-RAND с лучшим из EM)		среднее	+0,1164%
		σ	+95,5276%
Гипотеза о неравенстве мат. ожиданий $H: f^*_{GA\text{-}RAND} > f^*_{лучший\ EM}$		Статистика Стьюдента	5,69
		Порог статистики при α=1%	2,66
		Принятие гипотезы	да, преимущество статистически значимо

Примечание: * -лучший результат семейства алгоритмов (EM, GA, VNS), **- абсолютно лучший результат среди всех алгоритмов.

Таблица А.8 – сравнительные результаты работы VNS с жадной эвристикой для задач кластеризации и размещения.

Характеристика набора данных, задачи и алгоритма			Значение
Наименование набора данных			Europe (UCI), некорр.
Число векторов данных			N= 169308
Размерность пространства			d= 2
Время работы алгоритмов			t= 240 мин.
Число запусков алгоритмов			30
Число кластеров			k= 30
Алгоритм		Показатель	Значение
Модификации EM	EM	среднее	-3 653 148,61*
		σ	2 634,66
	CEM	среднее	-
		σ	-
	SEM	среднее	-3 654 493,58
		σ	4 348,55
	Лучший рез-т модификаций	среднее	-3 653 148,61
		σ	2 634,66
VNS-1		среднее	-3 648 660,14
		σ	1 383,53
(в сравнении VNS-1 с лучшим из EM)		среднее	+0,1229%
		σ	-47,4874%
Гипотеза о неравенстве мат. ожиданий $H:f^*_{VNS\text{-}1}>f^*_{лучший\ EM}$		Статистика Стьюдента	11,68
		Порог статистики при α=1%	2,66
		Принятие гипотезы	да, преимущество статистически значимо
VNS-2		среднее	-3 644 645,16
		σ	1 840,85
(в сравнении VNS-2 с лучшим из EM)		среднее	+0,2328%
		σ	-30,1294%
Гипотеза о неравенстве мат. ожиданий $H:f^*_{VNS\text{-}2}>f^*_{лучший\ EM}$		Статистика Стьюдента	20,49
		Порог статистики при α=1%	2,66
		Принятие гипотезы	да, преимущество статистически значимо
VNS-3		среднее	-3 639 340,87**
		σ	1 729,87
(в сравнении VNS-3 с лучшим из EM)		среднее	+0,3780%
		σ	-34,3418%
Гипотеза о неравенстве мат. ожиданий $H:f^*_{VNS\text{-}3}>f^*_{лучший\ EM}$		Статистика Стьюдента	33,93
		Порог статистики при α=1%	2,66
		Принятие гипотезы	да, преимущество статистически значимо

Примечание: * -лучший результат семейства алгоритмов (EM, GA, VNS), **- абсолютно лучший результат среди всех алгоритмов.

Таблица А.9 – сравнительные результаты работы ГА с жадной эвристикой для задач кластеризации и размещения.

Характеристика набора данных, задачи и алгоритма			Значение
Наименование набора данных			Europe (UCI), сферич.
Число векторов данных			N= 169308
Размерность пространства			d= 2
Время работы алгоритмов			t= 240 мин.
Число запусков алгоритмов			30
Число кластеров			k= 40
Алгоритм		Показатель	Значение
Модификации EM	EM	среднее	-3 625 957,39*
		σ	49,56
	CEM	среднее	-3 637 892,57
		σ	283,94
	SEM	среднее	-3 625 779,08
		σ	25,06
	Лучший рез-т модификаций	среднее	-3 625 779,08
		σ	25,06
GA-ONE		среднее	-3 625 878,21
		σ	36,34
(в сравнении GA-ONE с лучшим из EM)		среднее	-0,0027%
		σ	+44,9770%
Гипотеза о неравенстве мат. ожиданий $H: f^*_{GA\text{-}ONE} > f^*_{лучший\ EM}$		Статистика Стьюдента	-17,40
		Порог статистики при α=1%	2,66
		Принятие гипотезы	нет, преимущества нет
GA-FULL		среднее	-3 625 740,64*
		σ	15,19
(в сравнении GA-FULL с лучшим из EM)		среднее	+0,0011%
		σ	-39,4095%
Гипотеза о неравенстве мат. ожиданий $H: f^*_{GA\text{-}FULL} > f^*_{лучший\ EM}$		Статистика Стьюдента	10,16
		Порог статистики при α=1%	2,66
		Принятие гипотезы	да, преимущество статистически значимо
GA-RAND		среднее	-3 625 795,67
		σ	48,03
(в сравнении GA-RAND с лучшим из EM)		среднее	-0,0005%
		σ	+91,6177%
Гипотеза о неравенстве мат. ожиданий $H: f^*_{GA\text{-}RAND} > f^*_{лучший\ EM}$		Статистика Стьюдента	-2,37
		Порог статистики при α=1%	2,66
		Принятие гипотезы	нет, преимущества нет

Примечание: * -лучший результат семейства алгоритмов (EM, GA, VNS), **- абсолютно лучший результат среди всех алгоритмов.

Таблица А.10 – сравнительные результаты работы VNS с жадной эвристикой для задач кластеризации и размещения.

Характеристика набора данных, задачи и алгоритма			Значение
Наименование набора данных			Europe (UCI), сферич.
Число векторов данных			N= 169308
Размерность пространства			d= 2
Время работы алгоритмов			t= 240 мин.
Число запусков алгоритмов			30
Число кластеров			k= 40
Алгоритм		Показатель	Значение
Модификации EM	EM	среднее	-3 625 957,39*
		σ	49,56
	CEM	среднее	-3 637 892,57
		σ	283,94
	SEM	среднее	-3 625 779,08
		σ	25,06
	Лучший рез-т модификаций	среднее	-3 625 779,08
		σ	25,06
VNS-1		среднее	-3 625 691,71**
		σ	14,47
(в сравнении VNS-1 с лучшим из EM)		среднее	+0,0024%
		σ	-42,2676%
Гипотеза о неравенстве мат. ожиданий $H:f^*_{VNS\text{-}1}>f^*_{лучший\ EM}$		Статистика Стьюдента	23,39
		Порог статистики при α=1%	2,66
		Принятие гипотезы	да, преимущество статистически значимо
VNS-2		среднее	-3 625 694,18
		σ	20,15
(в сравнении VNS-2 с лучшим из EM)		среднее	+0,0023%
		σ	-19,6102%
Гипотеза о неравенстве мат. ожиданий $H:f^*_{VNS\text{-}2}>f^*_{лучший\ EM}$		Статистика Стьюдента	20,45
		Порог статистики при α=1%	2,66
		Принятие гипотезы	да, преимущество статистически значимо
VNS-3		среднее	-3 625 748,67
		σ	15,40
(в сравнении VNS-3 с лучшим из EM)		среднее	+0,0008%
		σ	-38,5468%
Гипотеза о неравенстве мат. ожиданий $H:f^*_{VNS\text{-}3}>f^*_{лучший\ EM}$		Статистика Стьюдента	8,01
		Порог статистики при α=1%	2,66
		Принятие гипотезы	да, преимущество статистически значимо

Примечание: * -лучший результат семейства алгоритмов (EM, GA, VNS), **- абсолютно лучший результат среди всех алгоритмов.

Таблица А.11 – сравнительные результаты работы ГА с жадной эвристикой для задач кластеризации и размещения.

Характеристика набора данных, задачи и алгоритма	Значение
Наименование набора данных	KDDCUP04Bio (UCI), нормир.
Число векторов данных	N= 145751
Размерность пространства	d= 74
Время работы алгоритмов	t= 500 мин.
Число запусков алгоритмов	30
Число кластеров	k= 30

Алгоритм		Показатель	Значение
Модификации EM	EM	среднее	-12 513 229,72
		σ	1 255,54
	CEM	среднее	-12 512 717,91*
		σ	343,03
	SEM	среднее	-12 514 336,57
		σ	683,61
	Лучший рез-т модификаций	среднее	-12 512 717,91
		σ	343,03
GA-ONE		среднее	-12 512 817,75
		σ	410,87
(в сравнении GA-ONE с лучшим из EM)		среднее	-0,0008%
		σ	+19,7792%
Гипотеза о неравенстве мат. ожиданий $H:f^*_{GA\text{-}ONE}>f^*_{лучший\ EM}$		Статистика Стьюдента	-1,44
		Порог статистики при α=1%	2,66
		Принятие гипотезы	нет, преимущества нет
GA-FULL		среднее	-12 512 517,03*
		σ	159,13
(в сравнении GA-FULL с лучшим из EM)		среднее	+0,0016%
		σ	-53,6100%
Гипотеза о неравенстве мат. ожиданий $H:f^*_{GA\text{-}FULL}>f^*_{лучший\ EM}$		Статистика Стьюдента	4,11
		Порог статистики при α=1%	2,66
		Принятие гипотезы	да, преимущество статистически значимо
GA-RAND		среднее	-12 512 811,47
		σ	639,09
(в сравнении GA-RAND с лучшим из EM)		среднее	-0,0007%
		σ	+86,3076%
Гипотеза о неравенстве мат. ожиданий $H:f^*_{GA\text{-}RAND}>f^*_{лучший\ EM}$		Статистика Стьюдента	-1,00
		Порог статистики при α=1%	2,66
		Принятие гипотезы	нет, преимущества нет

Примечание: * -лучший результат семейства алгоритмов (EM, GA, VNS), **- абсолютно лучший результат среди всех алгоритмов.

Таблица А.12 – сравнительные результаты работы VNS с жадной эвристикой для задач кластеризации и размещения.

Характеристика набора данных, задачи и алгоритма			Значение
Наименование набора данных			KDDCUP04Bio (UCI), нормир.
Число векторов данных			N= 145751
Размерность пространства			d= 74
Время работы алгоритмов			t= 500 мин.
Число запусков алгоритмов			30
Число кластеров			k= 30
Алгоритм		Показатель	Значение
Модификации EM	EM	среднее	-12 513 229,72
		σ	1 255,54
	CEM	среднее	-12 512 717,91*
		σ	343,03
	SEM	среднее	-12 514 336,57
		σ	683,61
	Лучший рез-т модификаций	среднее	-12 512 717,91
		σ	343,03
VNS-1		среднее	-12 511 984,94
		σ	278,04
(в сравнении VNS-1 с лучшим из EM)		среднее	+0,0059%
		σ	-18,9464%
Гипотеза о неравенстве мат. ожиданий $H: f^*_{VNS\text{-}1} > f^*_{лучший\ EM}$		Статистика Стьюдента	12,86
		Порог статистики при α=1%	2,66
		Принятие гипотезы	да, преимущество статистически значимо
VNS-2		среднее	-12 511 968,20**
		σ	105,26
(в сравнении VNS-2 с лучшим из EM)		среднее	+0,0060%
		σ	-69,3157%
Гипотеза о неравенстве мат. ожиданий $H: f^*_{VNS\text{-}2} > f^*_{лучший\ EM}$		Статистика Стьюдента	16,18
		Порог статистики при α=1%	2,66
		Принятие гипотезы	да, преимущество статистически значимо
VNS-3		среднее	-12 512 654,06
		σ	139,04
(в сравнении VNS-3 с лучшим из EM)		среднее	+0,0005%
		σ	-59,4667%
Гипотеза о неравенстве мат. ожиданий $H: f^*_{VNS\text{-}3} > f^*_{лучший\ EM}$		Статистика Стьюдента	1,34
		Порог статистики при α=1%	2,66
		Принятие гипотезы	нет, преимущество статистически не значимо

Примечание: * -лучший результат семейства алгоритмов (EM, GA, VNS), **- абсолютно лучший результат среди всех алгоритмов.

Таблица А.13 – сравнительные результаты работы ГА с жадной эвристикой для задач кластеризации и размещения.

Характеристика набора данных, задачи и алгоритма			Значение
Наименование набора данных			Missamerica (UCI)
Число векторов данных			N= 6480
Размерность пространства			d= 16
Время работы алгоритмов			t= 15 мин.
Число запусков алгоритмов			30
Число кластеров			k= 30
Алгоритм		Показатель	Значение
Модификации EM	EM	среднее	-261 971,11
		σ	99,57
	CEM	среднее	-262 265,49
		σ	51,64
	SEM	среднее	-261 839,62*
		σ	26,86
	Лучший рез-т модификаций	среднее	-261 839,62
		σ	26,86
GA-ONE		среднее	-261 893,52
		σ	71,72
(в сравнении GA-ONE с лучшим из EM)		среднее	-0,0206%
		σ	+167,0357%
Гипотеза о неравенстве мат. ожиданий $H: f^*_{GA\text{-}ONE} > f^*_{лучший\ EM}$		Статистика Стьюдента	-5,45
		Порог статистики при α=1%	2,66
		Принятие гипотезы	нет, преимущества нет
GA-FULL		среднее	-261 736,56*
		σ	1,62
(в сравнении GA-FULL с лучшим из EM)		среднее	+0,0394%
		σ	-93,9538%
Гипотеза о неравенстве мат. ожиданий $H: f^*_{GA\text{-}FULL} > f^*_{лучший\ EM}$		Статистика Стьюдента	29,67
		Порог статистики при α=1%	2,66
		Принятие гипотезы	да, преимущество статистически значимо
GA-RAND		среднее	-261 763,15
		σ	22,27
(в сравнении GA-RAND с лучшим из EM)		среднее	+0,0292%
		σ	-17,0860%
Гипотеза о неравенстве мат. ожиданий $H: f^*_{GA\text{-}RAND} > f^*_{лучший\ EM}$		Статистика Стьюдента	16,98
		Порог статистики при α=1%	2,66
		Принятие гипотезы	да, преимущество статистически значимо

Примечание: * -лучший результат семейства алгоритмов (EM, GA, VNS), **- абсолютно лучший результат среди всех алгоритмов.

Таблица А.14 – сравнительные результаты работы VNS с жадной эвристикой для задач кластеризации и размещения.

Характеристика набора данных, задачи и алгоритма			Значение
Наименование набора данных			Missamerica (UCI)
Число векторов данных			N= 6480
Размерность пространства			d= 16
Время работы алгоритмов			t= 15 мин.
Число запусков алгоритмов			30
Число кластеров			k= 30
Алгоритм		Показатель	Значение
Модификации EM	EM	среднее	-261 971,11
		σ	99,57
	CEM	среднее	-262 265,49
		σ	51,64
	SEM	среднее	-261 839,62*
		σ	26,86
	Лучший рез-т модификаций	среднее	-261 839,62
		σ	26,86
VNS-1		среднее	-261 741,88
		σ	10,04
(в сравнении VNS-1 с лучшим из EM)		среднее	+0,0373%
		σ	-62,6075%
Гипотеза о неравенстве мат. ожиданий $H:f^*_{VNS\text{-}1}>f^*_{лучший\ EM}$		Статистика Стьюдента	26,40
		Порог статистики при α=1%	2,66
		Принятие гипотезы	да, преимущество статистически значимо
VNS-2		среднее	-261 737,89
		σ	8,71
(в сравнении VNS-2 с лучшим из EM)		среднее	+0,0388%
		σ	-67,5674%
Гипотеза о неравенстве мат. ожиданий $H:f^*_{VNS\text{-}2}>f^*_{лучший\ EM}$		Статистика Стьюдента	27,91
		Порог статистики при α=1%	2,66
		Принятие гипотезы	да, преимущество статистически значимо
VNS-3		среднее	-261 737,34**
		σ	1,54
(в сравнении VNS-3 с лучшим из EM)		среднее	+0,0391%
		σ	-94,2510%
Гипотеза о неравенстве мат. ожиданий $H:f^*_{VNS\text{-}3}>f^*_{лучший\ EM}$		Статистика Стьюдента	29,45
		Порог статистики при α=1%	2,66
		Принятие гипотезы	да, преимущество статистически значимо

Примечание: * -лучший результат семейства алгоритмов (EM, GA, VNS), **- абсолютно лучший результат среди всех алгоритмов.

Таблица А.15 – сравнительные результаты работы ГА с жадной эвристикой для задач кластеризации и размещения.

Характеристика набора данных, задачи и алгоритма			Значение
Наименование набора данных			BIRCH-3 (UCI), сфер.
Число векторов данных			N= 100 000
Размерность пространства			d= 2
Время работы алгоритмов			t= 60 мин.
Число запусков алгоритмов			30
Число кластеров			k= 10
Алгоритм		Показатель	Значение
Модификации EM	EM	среднее	-2 704 718,95
		σ	4 738,75
	CEM	среднее	-2 693 183,37*
		σ	3 890,76
	SEM	среднее	-2 694 786,13
		σ	2 065,64
	Лучший рез-т модификаций	среднее	-2 693 183,37
		σ	3 890,76
GA-ONE		среднее	-2 694 222,34
		σ	1 451,49
(в сравнении GA-ONE с лучшим из EM)		среднее	-0,0386%
		σ	-62,6939%
Гипотеза о неравенстве мат. ожиданий $H: f^*_{GA\text{-}ONE} > f^*_{лучший\ EM}$		Статистика Стьюдента	-1,94
		Порог статистики при α=1%	2,66
		Принятие гипотезы	нет, преимущества нет
GA-FULL		среднее	-2 687 924,32**
		σ	1 169,86
(в сравнении GA-FULL с лучшим из EM)		среднее	+0,1953%
		σ	-69,9324%
Гипотеза о неравенстве мат. ожиданий $H: f^*_{GA\text{-}FULL} > f^*_{лучший\ EM}$		Статистика Стьюдента	10,03
		Порог статистики при α=1%	2,66
		Принятие гипотезы	да, преимущество статистически значимо
GA-RAND		среднее	-2 691 114,54
		σ	4 797,24
(в сравнении GA-RAND с лучшим из EM)		среднее	+0,0768%
		σ	+23,2982%
Гипотеза о неравенстве мат. ожиданий $H: f^*_{GA\text{-}RAND} > f^*_{лучший\ EM}$		Статистика Стьюдента	2,59
		Порог статистики при α=1%	2,66
		Принятие гипотезы	нет, преимущество статистически не значимо

Примечание: * -лучший результат семейства алгоритмов (EM, GA, VNS), **- абсолютно лучший результат среди всех алгоритмов.

Таблица А.16 – сравнительные результаты работы VNS с жадной эвристикой для задач кластеризации и размещения.

Характеристика набора данных, задачи и алгоритма			Значение
Наименование набора данных			BIRCH-3 (UCI), сфер.
Число векторов данных			N= 100 000
Размерность пространства			d= 2
Время работы алгоритмов			t= 60 мин.
Число запусков алгоритмов			30
Число кластеров			k= 10
Алгоритм		Показатель	Значение
Модификации EM	EM	среднее	-2 704 718,95
		σ	4 738,75
	CEM	среднее	-2 693 183,37*
		σ	3 890,76
	SEM	среднее	-2 694 786,13
		σ	2 065,64
	Лучший рез-т модификаций	среднее	-2 693 183,37
		σ	3 890,76
VNS-1		среднее	-2 701 230,12
		σ	5 567,43
(в сравнении VNS-1 с лучшим из EM)		среднее	-0,2988%
		σ	+43,0937%
Гипотеза о неравенстве мат. ожиданий $H: f^*_{VNS\text{-}1} > f^*_{лучший\ EM}$		Статистика Стьюдента	-9,18
		Порог статистики при α=1%	2,66
		Принятие гипотезы	нет, преимущества нет
VNS-2		среднее	-2 700 481,81
		σ	5 589,44
(в сравнении VNS-2 с лучшим из EM)		среднее	-0,2710%
		σ	+43,6594%
Гипотеза о неравенстве мат. ожиданий $H: f^*_{VNS\text{-}2} > f^*_{лучший\ EM}$		Статистика Стьюдента	-8,30
		Порог статистики при α=1%	2,66
		Принятие гипотезы	нет, преимущества нет
VNS-3		среднее	-2 688 604,40*
		σ	0,01
(в сравнении VNS-3 с лучшим из EM)		среднее	+0,1700%
		σ	-99,9996%
Гипотеза о неравенстве мат. ожиданий $H: f^*_{VNS\text{-}3} > f^*_{лучший\ EM}$		Статистика Стьюдента	9,12
		Порог статистики при α=1%	2,66
		Принятие гипотезы	да, преимущество статистически значимо

Примечание: * -лучший результат семейства алгоритмов (EM, GA, VNS), **- абсолютно лучший результат среди всех алгоритмов.

Таблица А.17 – сравнительные результаты работы ГА с жадной эвристикой для задач кластеризации и размещения.

Характеристика набора данных, задачи и алгоритма			Значение
Наименование набора данных			ИС 1526ИЕ10, некор.
Число векторов данных			N= 3987
Размерность пространства			d= 206
Время работы алгоритмов			t= 10 мин.
Число запусков алгоритмов			30
Число кластеров			k= 15
Алгоритм		Показатель	Значение
Модификации EM	EM	среднее	17 730 891,48*
		σ	867 120,28
	CEM	среднее	1 937 130,22
		σ	22 034,47
	SEM	среднее	986 494,89
		σ	397 882,68
	Лучший рез-т модификаций	среднее	17 730 891,48**
		σ	867 120,28
GA-ONE		среднее	8 840 784,77
		σ	5 206 942,01
(в сравнении GA-ONE с лучшим из EM)		среднее	-50,1391%
		σ	+500,4867%
Гипотеза о неравенстве мат. ожиданий $H: f^*_{GA\text{-}ONE} > f^*_{лучший\ EM}$		Статистика Стьюдента	-13,05
		Порог статистики при α=1%	2,66
		Принятие гипотезы	нет, преимущества нет
GA-FULL		среднее	12 438 569,93*
		σ	3 814 686,07
(в сравнении GA-FULL с лучшим из EM)		среднее	-29,8480%
		σ	+339,9258%
Гипотеза о неравенстве мат. ожиданий $H: f^*_{GA\text{-}FULL} > f^*_{лучший\ EM}$		Статистика Стьюдента	-10,48
		Порог статистики при α=1%	2,66
		Принятие гипотезы	нет, преимущества нет
GA-RAND		среднее	11 776 481,62
		σ	6 168 089,69
(в сравнении GA-RAND с лучшим из EM)		среднее	-33,5821%
		σ	+611,3303%
Гипотеза о неравенстве мат. ожиданий $H: f^*_{GA\text{-}RAND} > f^*_{лучший\ EM}$		Статистика Стьюдента	-7,40
		Порог статистики при α=1%	2,66
		Принятие гипотезы	нет, преимущества нет

Примечание: * -лучший результат семейства алгоритмов (EM, GA, VNS), **- абсолютно лучший результат среди всех алгоритмов.

Таблица А.18 – сравнительные результаты работы VNS с жадной эвристикой для задач кластеризации и размещения.

Характеристика набора данных, задачи и алгоритма			Значение
Наименование набора данных			ИС 1526ИЕ10, некор.
Число векторов данных			N= 3987
Размерность пространства			d= 206
Время работы алгоритмов			t= 10 мин.
Число запусков алгоритмов			30
Число кластеров			k= 15
Алгоритм		Показатель	Значение
Модификации EM	EM	среднее	17 730 891,48*
		σ	867 120,28
	CEM	среднее	1 937 130,22
		σ	22 034,47
	SEM	среднее	986 494,89
		σ	397 882,68
	Лучший рез-т модификаций	среднее	17 730 891,48**
		σ	867 120,28
VNS-1		среднее	5 090 536,31
		σ	2 316 820,27
(в сравнении VNS-1 с лучшим из EM)		среднее	-71,2900%
		σ	+167,1856%
Гипотеза о неравенстве мат. ожиданий $H:f^*_{VNS\text{-}1}>f^*_{лучший\ EM}$		Статистика Стьюдента	-39,58
		Порог статистики при α=1%	2,66
		Принятие гипотезы	нет, преимущества нет
VNS-2		среднее	9 824 827,99
		σ	5 216 417,51
(в сравнении VNS-2 с лучшим из EM)		среднее	-44,5892%
		σ	+501,5795%
Гипотеза о неравенстве мат. ожиданий $H:f^*_{VNS\text{-}2}>f^*_{лучший\ EM}$		Статистика Стьюдента	-11,58
		Порог статистики при α=1%	2,66
		Принятие гипотезы	нет, преимущества нет
VNS-3		среднее	14 215 864,72*
		σ	4 401 058,87
(в сравнении VNS-3 с лучшим из EM)		среднее	-19,8243%
		σ	+407,5488%
Гипотеза о неравенстве мат. ожиданий $H:f^*_{VNS\text{-}3}>f^*_{лучший\ EM}$		Статистика Стьюдента	-6,07
		Порог статистики при α=1%	2,66
		Принятие гипотезы	нет, преимущества нет

Примечание: * -лучший результат семейства алгоритмов (EM, GA, VNS), **- абсолютно лучший результат среди всех алгоритмов.

Таблица А.19 – сравнительные результаты работы ГА с жадной эвристикой для задач кластеризации и размещения.

Характеристика набора данных, задачи и алгоритма			Значение
Наименование набора данных			ИС140УД25АС1ВК, некор.
Число векторов данных			N= 56
Размерность пространства			d= 43
Время работы алгоритмов			t= 2 мин.
Число запусков алгоритмов			30
Число кластеров			k= 10
Алгоритм		Показатель	Значение
Модификации EM	EM	среднее	7 541,08*
		σ	457,95
	CEM	среднее	853,61
		σ	7,88
	SEM	среднее	928,89
		σ	105,52
	Лучший рез-т модификаций	среднее	7 541,08
		σ	457,95
GA-ONE		среднее	7 291,67
		σ	473,67
(в сравнении GA-ONE с лучшим из EM)		среднее	-3,3074%
		σ	+3,4336%
Гипотеза о неравенстве мат. ожиданий $H: f^*_{GA\text{-}ONE} > f^*_{лучший\ EM}$		Статистика Стьюдента	-2,93
		Порог статистики при α=1%	2,66
		Принятие гипотезы	нет, преимущества нет
GA-FULL		среднее	7 692,40*
		σ	1 335,78
(в сравнении GA-FULL с лучшим из EM)		среднее	+2,0065%
		σ	+191,6895%
Гипотеза о неравенстве мат. ожиданий $H: f^*_{GA\text{-}FULL} > f^*_{лучший\ EM}$		Статистика Стьюдента	0,83
		Порог статистики при α=1%	2,66
		Принятие гипотезы	нет, преимущество статистически не значимо
GA-RAND		среднее	7 551,50
		σ	438,48
(в сравнении GA-RAND с лучшим из EM)		среднее	+0,1381%
		σ	-4,2520%
Гипотеза о неравенстве мат. ожиданий $H: f^*_{GA\text{-}RAND} > f^*_{лучший\ EM}$		Статистика Стьюдента	0,13
		Порог статистики при α=1%	2,66
		Принятие гипотезы	нет, преимущество статистически не значимо

Примечание: * -лучший результат семейства алгоритмов (EM, GA, VNS), **- абсолютно лучший результат среди всех алгоритмов.

Таблица А.20 – сравнительные результаты работы VNS с жадной эвристикой для задач кластеризации и размещения.

Характеристика набора данных, задачи и алгоритма			Значение
Наименование набора данных			ИС140УД25АС1ВК, некор.
Число векторов данных			N= 56
Размерность пространства			d= 43
Время работы алгоритмов			t= 2 мин.
Число запусков алгоритмов			30
Число кластеров			k= 10
Алгоритм		Показатель	Значение
Модификации EM	EM	среднее	7 541,08*
		σ	457,95
	CEM	среднее	853,61
		σ	7,88
	SEM	среднее	928,89
		σ	105,52
	Лучший рез-т модификаций	среднее	7 541,08
		σ	457,95
VNS-1		среднее	7 797,86**
		σ	1 129,79
(в сравнении VNS-1 с лучшим из EM)		среднее	+3,4051%
		σ	+146,7064%
Гипотеза о неравенстве мат. ожиданий $H: f^*_{VNS\text{-}1} > f^*_{лучший\ EM}$		Статистика Стьюдента	1,63
		Порог статистики при α=1%	2,66
		Принятие гипотезы	нет, преимущество статистически не значимо
VNS-2		среднее	7 267,67
		σ	1 013,48
(в сравнении VNS-2 с лучшим из EM)		среднее	-3,6256%
		σ	+121,3104%
Гипотеза о неравенстве мат. ожиданий $H: f^*_{VNS\text{-}2} > f^*_{лучший\ EM}$		Статистика Стьюдента	-1,90
		Порог статистики при α=1%	2,66
		Принятие гипотезы	нет, преимущества нет
VNS-3		среднее	7 138,46
		σ	549,37
(в сравнении VNS-3 с лучшим из EM)		среднее	-5,3390%
		σ	+19,9632%
Гипотеза о неравенстве мат. ожиданий $H: f^*_{VNS\text{-}3} > f^*_{лучший\ EM}$		Статистика Стьюдента	-4,36
		Порог статистики при α=1%	2,66
		Принятие гипотезы	нет, преимущества нет

Примечание: * -лучший результат семейства алгоритмов (EM, GA, VNS), **- абсолютно лучший результат среди всех алгоритмов.

ПРИЛОЖЕНИЕ Б. Акт об использовании результатов исследования

УТВЕРЖДАЮ
Директор
ОАО «Испытательный технический центр -
НПО ПМ» канд. техн. наук
В.И. Орлов

АКТ
о внедрении результатов диссертационного исследования
Сташкова Дмитрия Викторовича

Настоящим актом подтверждается, что алгоритмы метода жадных эвристик для решения задач автоматической группировки объектов на основе разделения смеси распределений внедрены в опытную эксплуатацию в составе системы автоматизированного формирования и контроля специальных партий электрорадиоизделий космического применения на ОАО «Испытательный технический центр – НПО ПМ» (г.Железногорск).

Применение комбинации алгоритмов, а также модели разделения сборных партий промышленной продукции на однородные партии на основе модели разделения смеси распределений для систем автоматической группировки объектов, разработанных в рамках диссертационного исследования соискателя ученой степени кандидата технических наук Сташкова Дмитрия Викторовича, позволило снизить процент ошибок при выявлении однородных партий электрорадиоизделий.

Зам директора «ОАО ИТЦ –НПО ПМ»
канд. техн. наук
В.В. Федосов