Санкт-Петербургский государственный университет

ЛОЖКИНС Алексейс

Выпускная квалификационная работа

«Исследование робастности кластеризации методами статистического и имитационного моделирования»

Уровень образования:

Направление 02.06.01 «Компьютерные и информационные науки»

Основная образовательная программа МК.3005.2016 «Математическая кибернетика»

Научный руководитель:
д. т. н., проф. кафедры
математической теории игр
и статистических решений,
В. М. Буре

Рецензент:
д. ф.-м. н., проф., зав.
кафедры компьютерной
инженерии и программотех-
ники РГПУ им. А. И. Герцена,
А. В. Флегонтов

Санкт-Петербург
2019 г.

Содержание

Введение

В первой половине XX века кластерный анализ начинает свое развитие как отдельное от таксономии направление. Польским антропологом Яном Чекановски в 1911 году была опубликована одна из первых работ в области кластерного анализа [1]. В работе выдвигается идея о формировании групп близких элементов, что составляет основную цель кластерного анализа как инструмента анализа данных.

Развитие компьютерных технологий и вычислительных инструментов предоставляет возможность, а увеличение хранимой информации создает спрос на методы и подходы обработки данных. Кластерный анализ — одна из задач обучения без учителя — подразумевает, что по некоторому правилу определяется схожесть объектов и наиболее близкие объекты образуют группы или кластеры. Природа анализируемых данных может значительно отличаться в зависимости от решаемой задачи, области применения и набора параметров. Это приводит к тому, что нет универсального метода кластеризации или процедуры исследования, а кластерный анализ — это набор методов и алгоритмов для решения задач анализа данных, который непрерывно развивается.

Понятие «кластер» не имеет строгого математического описания, что предоставляет исследователям возможность разрабатывать методы и алгоритмы решения задач кластеризации, ориентируясь на специфику анализируемых данных, и предлагать критерии по оценке качества кластеризации. Можно выделить несколько задач кластерного анализа:

1. Выбор методов кластеризации. Алгоритмы формирования разбиения данных на группы, где объекты из одной группы близки, а объекты из разных групп отличаются.

2. Оценка качества кластеризации. Критерии оценки методов кластери-

зации для сравнения и выявления наиболее «правильной».

3. Оценка количества кластеров. Методы определения количества групп разбиения.

4. Выбор функции расстояния. Способ сравнения объектов и оценка их близости.

Обзор методов кластеризации, критериев качества кластеризации и функций расстояния представлен в работах [2; 3].

В литературе [4] выделяют несколько целей кластеризации:

1. Сокращение объемов хранимых данных. Фиксирование одного представителя каждого кластера и удаление остальных. Используется в сжатии данных или для сокращения объемов исследуемых данных.

2. Определение объектов, не принадлежащих ни к одному кластеру. Задачи одноклассовой классификации или обнаружение новизны (англ. novelty detection).

3. Построение иерархического множества объектов (задачи таксономии).

4. Разбиение задачи обработки данных. Анализ данных внутри кластера вместо анализа всего набора данных.

В выпускной квалификационной работе (ВКР) исследуется одна из задач кластерного анализа — определение «правильной» кластеризации и, как следствие, определение «правильного» количества кластеров. В общем случае выбор количества кластеров зависит не только от методов кластеризации и функций расстояния между объектами, но и от методов сравнения результатов кластеризации. В настоящей работе предлагается концепция решения задачи о нахождении качественной кластеризации, основанная на стохастической устойчивости кластеризации. Применяются процедуры статистического и имитационного моделирования для реализации возможных

изменений исследуемых данных или учета возможных ошибок (ошибки измерения, ошибки в представлении чисел в ЭВМ, шум, неполнота данных и т.п.), сравнения кластеризаций и определение уровня рисков отклонения кластеризаций.

Предметом исследования является кластерный анализ, а объектом исследования — алгоритмы определения устойчивой кластеризации.

Цель ВКР заключается в разработке критерия определения статистически устойчивой (надежной) кластеризации с использованием процедур статистического и имитационного моделирования.

Достижение поставленной цели требует решения следующих **задач**:

1. Задача разработки статистической процедуры анализа кластеризаций.

2. Задача повышения эффективности статистических процедур.

3. Задача оценки и выбора статистически устойчивой кластеризации.

Теоретическая новизна заключается в следующем:

1. Разработана процедура возмущения кластеризации на основе имитационного моделирования.

2. Разработаны 2 критерия определения устойчивой кластеризации основанных на теории рисков и математическом ожидании сменяемости кластеризации.

3. Разработана метрика сравнения близости кластеризаций.

Практическая значимость работы заключается в следующем:

1. Разработана программа ЭВМ для численного решения задачи нахождения устойчивого размещения хабов на основе процедур введенного критерия определения устойчивой кластеризации.

2. Разработана программа ЭВМ для определения устойчивой кластеризации на основе разработанных критериев. Проведено сравнение существующих критериев нахождения устойчивой кластеризации с предлагаемыми в настоящей ВКР.

Основные результаты ВКР были представлены и обсуждены на следующих конференциях:

1. XLVI Международная научная конференция аспирантов и студентов «Процессы управления и устойчивость», 6—9 апреля 2015 г., г. Санкт-Петербург;

2. 20th International Conference on Mathematical Modelling and Analysis, May 26–29, 2015, Sigulda, Latvia;

3. III Международная конференция «Устойчивость и процессы управления», посвященная 85-летию со дня рождения профессора, чл.-корр. РАН В. И. Зубова, 5—9 октября 2015 г., г. Санкт-Петербург;

4. XLIX Международная научная конференция аспирантов и студентов «Процессы управления и устойчивость», 2—5 апреля 2018 г., г. Санкт-Петербург;

5. XIV Международная научная конференция «Устойчивость и колебания нелинейных систем управления» (конференция Пятницкого), 30 мая — 1 июня 2018 г., г. Москва.

Публикации. Материалы диссертации опубликованы в 14 печатных работах, из них 6 тезисов докладов [5—10], 2 статьи в трудах конференций [11; 12], индексируемых в РИНЦ, 1 статья в трудах конференции, индексируемом в библиографических базах данных Scopus и Web of Science [13], 1 статья в журнале РИНЦ [14], 1 статья в журнале, индексируемом в базе

Web of Science [15], и 2 статьи в журналах, индексируемых в базах Scopus и Web of Science [16; 17], 1 работа является магистерской диссертацией автора [18]. Получено свидетельство о государственной регистрации программы для ЭВМ [19].

Структура ВКР. ВКР состоит из введения, постановки задачи, обзора литературы, трех глав, выводов, заключения и списка литературы.

Во Введении отражена актуальность работы, поставлена цель исследования, обоснована научная новизна работы, кратко описаны полученные результаты, и показана их практическая ценность.

В постановке задачи введены основные обозначения и понятия, сформулирована задача кластерного анализа, зафиксирована постановка задачи, которая решается в настоящей работе.

В обзоре литературы приведена классификация методов и подходов к решению поставленной задачи, обсуждены основные принципы работы методов из каждого класса.

В Главе 1 представлены формулировка двух критериев устойчивой кластеризации и процедура их расчета.

В Главе 2 представлены результаты численного эксперимента на искусственных данных и реальных данных, проведено сравнение результатов с другими индексами устойчивости.

В Главе 3 представлены результаты применения критериев устойчивости к задаче о размещении хабов в сети, проведен численный эксперимент на реальных данных.

В заключении подведены итоги работы и сформулированы основные выводы.

Общий объем ВКР составляет 57 страниц, включая 3 рисунка и 10 таблиц. Список литературы включает 59 наименований.

Постановка задачи

Кластерный анализ является многомерной статистической процедурой, где объекты исследования, $x_i \in X \subset R^m$, представлены в виде m-мерного вектора. Размерность m определяет количество исследуемых признаков объекта. Множество номеров/меток кластеров обозначим Y. Конечная выборка из n объектов $X^{[n]} \subset X$, тогда задача кластеризации состоит в разбиении множества объектов $X^{[n]}$ на непересекающиеся группы, называемые кластерами. Объекты внутри одной группы ближе по заранее выбранной метрике ρ, чем объекты из разных кластеров.

Алгоритм кластеризации $\alpha : X \to Y$ для каждого объекта $x_i \in X^{[n]}$ ставит в соответствие метку кластера $y_i \in Y$, обзор методов кластеризации приведен в работах [2; 3]. В настоящей работе используется матричное представление кластеризации, где элемент матрицы:

$$c_{ij} = \begin{cases} 1, & \text{если } x_i \text{ и } x_j \text{ принадлежат одному кластеру и } x_i \neq x_j; \\ 0, & \text{в противном случае.} \end{cases} \tag{1}$$

Размерность матрицы результата кластеризации $n \times n$, и из определения матрицы следует ее симметричность.

В кластерном анализе методы перекрестной проверки для определения числа кластеров предлагают функции от матриц кластеризации для получения промежуточных характеристик или конечных результатов. Введем операцию скалярного произведения двух матриц кластеризации:

$$\langle C^{k_1}, C^{k_2} \rangle = \sum_{i,j} c_{ij}^{k_1} c_{ij}^{k_2},$$

где C^{k_1} и C^{k_2} представляют два результата кластеризации для одинакового или разного количества кластеров, полученных на $X^{[n]}$. Результатом скалярного произведения $\langle C^{k_1}, C^{k_2} \rangle$ является количество пар объектов, лежащих в одном кластере одновременно в двух кластеризациях.

Скалярное произведение удовлетворяет неравенству Коши–Шварца [20]:

$$\langle C^{k_1}, C^{k_2} \rangle \leq \sqrt{\langle C^{k_1}, C^{k_1} \rangle \langle C^{k_2}, C^{k_2} \rangle}.$$

Элементы матрицы C обладают свойством транзитивности: если $c_{ij} = 1$ и $c_{jk} = 1$, то $c_{ik} = 1$ и $c_{ki} = 1$.

В настоящей ВКР автором предлагается подход к определению числа кластеров, основанный на статистической устойчивости кластеров. Устойчивость кластеризации определяется на основе сравнения кластеризаций между собой и оценки уровня «близости» разбиений. Идея сравнения кластеризаций не является новой. Ниже представлены некоторые функции сравнения кластеризаций, которые могут быть использованы в предлагаемом подходе к оценке количества кластеров.

В работе [21] для сравнения кластеризаций используется коэффициент Отиаи [22]. Данный коэффициент широко используется в глубинном обучении для обработки естественного языка, машинного перевода и в кластерном анализе. Коэффициант Отиаи в введенных обозначениях имеет следующий вид:

$$cosine(C^{k_1}, C^{k_2}) = \frac{\langle C^{k_1}, C^{k_2} \rangle}{\sqrt{\langle C^{k_1}, C^{k_1} \rangle \langle C^{k_2}, C^{k_2} \rangle}}. \quad (2)$$

Данный коэффициент представляет нормализованную корреляцию между двумя кластеризациями, которая может быть рассмотрена в качестве меры близости двух кластеризаций. В случае, когда матрицы $C^{k_1} = C^{k_2}$, Коэффициант Отиаи будет равен 1. Как в случае с коэффициентом корреляции, чем ближе значение коэффицианта Отиаи к 1, тем сильнее сходство кластеризаций.

Коэффициент Жаккара [23] используется для сравнения кластеризаций, где для оценки учитываются только совпадения в результатах, а расхождения не влияют на результат. В введенных обозначениях коэффициент

Жаккара будет иметь вид:

$$J(C^{k_1}, C^{k_2}) = \frac{\langle C^{k_1}, C^{k_2} \rangle}{\langle C^{k_1}, C^{k_1} \rangle + \langle C^{k_2}, C^{k_2} \rangle - \langle C^{k_1}, C^{k_2} \rangle}. \quad (3)$$

В работе [24] предлагается коэффициент сравнения кластеризаций, основанный на взвешенном штрафе расхождений в результатах двух кластеризаций:

$$M(C^{k_1}, C^{k_2}) = 1 - \frac{1}{n^2} \parallel C^{k_1} - C^{k_2} \parallel^2, \quad (4)$$

где $\parallel C \parallel^2 = \langle C, C \rangle$ — норма матрицы.

Выбор функции сравнения кластеризаций зависит от целей исследования, природы данных, объема данных и других факторов.

Введем «грубую» функцию сравнения кластеризаций:

$$lb(C^{k_1}, C^{k_2}) = \begin{cases} 0, & \text{если } \frac{\langle C^{k_1}, C^{k_2} \rangle}{\sqrt{\langle C^{k_1}, C^{k_1} \rangle \langle C^{k_2}, C^{k_2} \rangle}} = 1; \\ 1, & \text{в противном случае.} \end{cases} \quad (5)$$

Представленная функция оценивает наличие полного совпадения кластеризаций. Если кластеризации совпадают, то уровень отличия кластеризаций равен 0. Оценка близости является грубой, независимо от количества отличий между кластеризациями, мера сходства будет указывать только на наличие или отсутствие отличий. Справедлива следующая лемма:

Лемма. Функция сравнения кластеризаций $lb(C^{k_1}, C^{k_2})$ является метрикой, когда соответствующие матрицы кластеризаций C^{k_1}, C^{k_2} представляют результаты разбиений одного и того же набора данных на одинаковое количество кластеров.

Доказательство. Функция сравнения кластеризаций удовлетворяет условию неотрицательности: $lb(C^{k_1}, C^{k_2}) \geq 0$, множество принимаемых значений $\{0, 1\}$.

Покажем, что рассматриваемая мера сходства удовлетворяет аксиоме тождества. Равенство $lb(C^{k_1}, C^{k_2}) = 0$ справедливо тогда и только тогда,

когда

$$\frac{\langle C^{k_1}, C^{k_2}\rangle}{\sqrt{\langle C^{k_1}, C^{k_1}\rangle\langle C^{k_2}, C^{k_2}\rangle}} = 1. \quad (6)$$

Если $C^{k_1} = C^{k_2}$, тогда $\langle C^{k_1}, C^{k_2}\rangle = \langle C^{k_1}, C^{k_1}\rangle = \langle C^{k_2}, C^{k_2}\rangle$, следовательно, выполняется (6).

Если $C^{k_1} \neq C^{k_2}$, тогда $\langle C^{k_1}, C^{k_2}\rangle < \langle C^{k_1}, C^{k_1}\rangle$ и $\langle C^{k_1}, C^{k_2}\rangle < \langle C^{k_2}, C^{k_2}\rangle$, так как $\langle C^{k_1}, C^{k_1}\rangle$ и $\langle C^{k_2}, C^{k_2}\rangle$ имеют максимальное количество совпадений (следует из определения матрицы C), следовательно, $\frac{\langle C^{k_1}, C^{k_2}\rangle}{\sqrt{\langle C^{k_1}, C^{k_1}\rangle\langle C^{k_2}, C^{k_2}\rangle}} < 1$ и $lb(C^{k_1}, C^{k_2}) = 1$.

Свойство симметричности наследуется из свойств скалярного произведения:
$lb(C^{k_1}, C^{k_2}) = lb(C^{k_2}, C^{k_1})$.

Покажем, что выполняется неравенство треугольника.

Если $lb(C^{k_1}, C^{k_2}) = 0$, следовательно, $lb(C^{k_1}, C^{k_2}) \leq lb(C^{k_1}, C^{k_3}) + lb(C^{k_2}, C^{k_3})$ выполняется, так как $lb(C^{k_1}, C^{k_3}) + lb(C^{k_2}, C^{k_3}) \geq 0$ вытекает из свойства неотрицательности.

Если $lb(C^{k_1}, C^{k_2}) = 1$, следовательно, $C^{k_1} \neq C^{k_2}$. Покажем, что $lb(C^{k_1}, C^{k_2}) \leq lb(C^{k_1}, C^{k_3}) + lb(C^{k_2}, C^{k_3})$. Без потери общности, пусть $C^{k_1} = C^{k_3}$. Следовательно, $C^{k_2} \neq C^{k_3}$. В этом случае $lb(C^{k_1}, C^{k_3}) + lb(C^{k_2}, C^{k_3}) = 1$, и неравенство треугольника выполняется.

Если $C^{k_1} \neq C^{k_3}$ и $C^{k_2} \neq C^{k_3}$, тогда $lb(C^{k_1}, C^{k_3}) + lb(C^{k_2}, C^{k_3}) = 2$, что влечет выполнение неравенстве треугольника. **Доказано.**

Введенная метрика позволяет сравнивать две кластеризации на наличие полного совпадения в разбиении на группы. Задача нахождения числа кластеров состоит в нахождении такой матрицы C, которая является «наилучшей» в сравнении с другими разбиениями для различных k. Метрики сравнения кластеризаций используются для определения различий между кластеризациями.

Формализация задачи. Требуется найти для выборки данных $X^{[n]}$

кластеризацию и количество кластеров, которые являются «истинными» или «правильными» (соответствуют истинному разбиению). Задача состоит в нахождении алгоритма кластеризации $\alpha(X^{[n]})$, функции расстояния между объектами и определении количества кластеров, которые приводят к требуемой кластеризации.

В силу того, что все алгоритмы кластеризации и функции расстояний между объектами представляют неограниченное количество комбинаций, то «лучшей» кластеризацией называется кластеризация, наиболее близкая к «правильной[1]» на множестве рассматриваемых алгоритмов кластеризации, функций расстояния между объектами и множества значений количества кластеров.

[1] Под «правильной» (англ. «natural/correct» clustering) кластеризацией здесь и в публикуемой литературе по тематике предполагается естественная кластеризация. Точного математического определения оптимальной кластеризации не существует, в противном случае, задача будет относится к разделу машинного обучения «Обучение с учителем».

Обзор литературы

Набор методов кластеризации можно разделить на концептуальные группы, где алгоритмы из одной группы используют схожую идею для кластеризации и интерпретации данных. Строгой общепринятой классификации методов кластеризации нет, но в литературе часто выделяют следующие подходы:

1. Иерархический подход. Методы, использующие функции расстояний для формирования вложенных групп (дендограмм).

2. Центроидная модель. Алгоритмы этой группы используют вероятностный подход для нахождения центров кластеров усреднением значений признаков. К таким методам относятся k-средних, k-медиан, EM-алгоритм, алгоритмы семейства FOREL и др.

3. Подходы, основанные на плотности данных. Алгоритмы DBSCAN и OPTICS определяют кластеры как области сгущения данных (уплотнения данных).

4. Графовые подходы. Представление данных в виде графа $G = (V, E)$, где вершинам соответствуют объекты, а ребра имеют вес, который является функцией близости объектов (например, функция расстояния). Метод связных компонент, алгоритм послойной кластеризации и алгоритм построения минимального покрывающего дерева являются представителями методов кластеризации, основанных на графах.

5. Подходы, основанные на системе искусственного интеллекта.

Широкий обзор методов кластеризации и одна из классификаций методов кластеризации представлены в [25].

Другое деление методов кластеризации — на четкие и нечеткие — основано на однозначности соотношения объектов и кластеров, где объект

может принадлежать строго одному кластеру или с некоторой вероятностью к каждому кластеру соответственно. Алгоритмы четкой и нечеткой кластеризации рассмотрены в работе [26].

Во введении говорится о задаче валидации числа кластеров как подзадачи кластерного анализа. В задаче количества кластеров, в литературе количество кластеров часто имеет обозначение k, как в алгоритме кластеризации k-средних, рассматривается фиксированное количество кластеризаций и ограниченный набор значений k, выбранный экспертом, что приводит к ситуации, в которой нельзя утверждать о получении глобальных и объективных результатов. Также увеличение значения k приводит к тому, что кластеры становятся более плотными и «качественными». В критерии определения числа групп этот фактор должен штрафоваться, чтобы избежать ситуации, когда количество кластеров совпадает с количеством объектов в наборе данных.

В работе [27] обозначены два вопроса, ответы на которые приведут к удовлетворительному решению проблемы:

1. Какое деление объектов на фиксированное количество групп k можно считать лучшим?

2. Какое значение k лучшее?

Ответы на эти вопросы целиком зависят от выбора целевой функции, оценивающей кластеры, и алгоритма оптимизации целевой функции. Второй вопрос подразумевает, что просматриваются возможные значения k и выбирается лучшее разбиение на группы. Критерий выбора лучшего разбиения должен быть связан с целевой функцией разбиения.

В литературе представлено множество методов решения задачи определения числа кластеров, которые будут рассмотрены ниже.

Индексные методы определения количества кластеров. Методы данной группы используют индексы для оценки качества кластеризации. В основе таких методов лежит соотношение степени разброса дан-

ных внутри кластеров к разбросу между кластерами. Критерий выбора количества кластеров целесообразно выбирать согласно целевой функции алгоритма кластеризации. Использование индексных методов определения количества кластеров подразумевает наличие предварительных кластеризаций для их сравнения между собой.

Один из первых индексных методов был предложен в работе [28]

$$k_{best} = argmax_{k \in K}(\frac{BGSS(k)}{k-1} / \frac{WGSS(k)}{n-k}),$$

где $BGSS(k)$ и $WGSS(k)$ — сумма квадратов расстояний между элементами разных кластеров (англ. between-group sum of squares) и внутри одного кластера (англ. within-group sum of squares), соответственно, K является множеством рассматриваемых значений k, а k_{best} — лучшее число кластеров k для кластеризации из рассмотренного множества K, n — количество объектов в исследуемой выборке данных. Данный критерий является аналогом F-статистики в дисперсионном анализе.

В работе [29] авторы отвечают на поставленные в [27] вопросы, предлагая следующий индексный метод оценки количества кластеров:

$$k_{best} = argmax_{k \in K} |\frac{DIFF(k)}{DIFF(k+1)}|,$$

где $DIFF(k) = (k-1)^{\frac{2}{p}} WGSS(k-1) - k^{\frac{2}{p}} WGSS(k)$, p — количество анализируемых признаков объектов.

Хартиган в своей работе [30] представляет еще один критерий:

$$k_{best} = min\{k | (n-k-1)[\frac{WGSS(k)}{WGSS(k+1)} - 1] \leq 10, \quad \forall k \in K\}.$$

К индексным методам относятся также индекс Davies-Bouldin [31], Dunn-индекс [32], Fowlkes–Mallows [21] и др. В работах [33; 34] значения индексов используются для построения аппроксимационных функций для анализа экстремальных точек.

Подход к оценке количества кластеров, основанный на информационных критериях. В случае, когда возможно построить функцию правдоподобия для модели кластеризации, для определения числа кластеров представляется возможным использовать информационные критерии, такие как AIC, BIC или DIC.

В работе [35] рассматривается следующая функция правдоподобия для выявления наиболее вероятной кластеризации:

$$P(D|M,\sigma^2) = \prod_{j\in D} P(u_j|M,\sigma^2),$$

где $P(u_j|M,\sigma^2) = \frac{1}{k}\sum_{i=1}^{k} \frac{1}{\sqrt{2\pi\sigma^2}} exp[-\frac{1}{2}\frac{(u_j-\mu_i)^2}{\sigma^2}]$ — сумма по k одинаково взвешенных нормальных распределений со средним значением μ_i и общей дисперсией σ^2 для объекта $u_j \in D \subset R^p$, а $M = \{\mu_i\}_{i=1}^k$ — множество центров кластеров. Общая дисперсия $\sigma^2 = \frac{1}{n}\sum_{i=1}^{k}\sum_{j\in C_i} ||u_j - \mu_i||^2$, где C_i — множество индексов элементов, принадлежащих кластеру i. Выбор количества кластеров осуществляется на основе максимизации информационного критерия:

Информационный критерий Акайке [36]:

$$k_{best} = argmax_{k\in K}[\ln P(D|M,\sigma^2) - (kp+1)].$$

Байесовский информационный критерий Шварца [37]:

$$k_{best} = argmax_{k\in K}[\ln P(D|M,\sigma^2) - \frac{(kp+1)}{2}\ln(n)].$$

В работе [38] авторами предлагается еще один информационный критерий:

$$k_{best} = argmax_{k\in K}[\ln P(D|M,\sigma^2) - \frac{(kp+1)}{2}\ln(n) -$$
$$-\sum_{s=1}^{n}\ln(s+\frac{k+2}{2}) + \sum_{i=1}^{k}\sum_{j=1}^{n_i}\ln(j+\frac{3}{2})],$$

где $|C_i| = n_i$ — количество элементов в кластере i и выполняется $\sum_{i=1}^{k} n_i = n$.

Информационно–теоретический подход к определению количества кластеров. В основе данного подхода лежит оценка внутрикластерной дисперсии и ее «скачка» (англ. jump) при переходе от одного количества кластеров к другому. В работе [39] применяется статистика «скачка», вводится функция расстояния:

$$d(j, \mu_i) = (u_j - \mu_i)^T \Gamma_i^1 (u_j - \mu_i),$$

где Γ_i^1 — внутрикластерная ковариационная матрица. «Скачек» определяется как $JS(k) = WGSS(k)^{-\frac{p}{2}} - WGSS(k-1)^{-\frac{p}{2}}$, предполагается, что $WGSS(0)^{-\frac{p}{2}} = 0$. Максимальный «скачек» соответствует «правильному» количеству кластеров.

Данный метод имеет следующее обоснование: если данные принадлежат нескольким нормальным распределениям, при которых расстояния между центрами распределений достаточно велики, тогда «скачек» должен произойти при k, равном числу нормальных распределений.

Метод «локтя» (англ. Elbow method). Визуальный метод определения числа кластеров — метод "локтя"[40] — является самым старым подходом к решению поставленной задачи. Вводится понятие функции стоимости $J(k)$, которая зависит от метода кластеризации, количества кластеров и других параметров. Число k_{best} считается «хорошим», если, начиная с $k = 2$ и при дальнейшем увеличении на каждом шаге на 1, существует значение k_{best}, при котором значение функции затрат резко падает, по сравнению со значением функции затрат для предыдущего числа кластеров, а затраты для последующих значений количества кластеров более пологие.

Основная проблема данного метода — это величина падения и понятие пологости графика.

Перекрестная проверка (англ. Cross-validation). В методах пе-

рекрестной проверки используется процедура разбиения всего множества данных на выборки случайным образом, которые в зависимости от алгоритма могут пересекаться или нет. На первом этапе фиксируется одна из выборок, в которой производится разбиение объектов выборки на группы. Данное разбиение служит эталоном (примером) и используется для сравнения с результатами кластеризации на других выборках. Кластеризация считается «хорошей», если в результате кластеризации каждой выборки получаются схожие результаты. Оценка схожести результатов на различных подмножествах и способы разбиения данных являются темой исследования в работах [20; 41; 42]:

Методы определения числа кластеров, основанные на устойчивости, также относятся к перекрестной проверке.

В работе [43] предложен подход, близкий описанному в настоящей диссертации в Главе 1. Алгоритм основан на выделении случайным образом пересекающихся подмножеств с количеством объектов, большим половины объектов в исходной выборке. Метод кластеризации применяется к каждому подмножеству, а результат кластеризации сравнивается на пересечении множеств. Идея основана на сохранении структуры данных на случайных подмножествах, следовательно, кластеризация подмножеств сохраняет структуру кластеров. В работе также упоминается добавление случайного шума в данные, вместо подмножества используется возмущенное исходное множество данных.

Задача оценки количества кластеров имеет пересечение с задачей оценки качества кластеризации. Эта связь позволяет использовать методы и подходы к решению одной задачи и применять к другой. Обзор методов оценки качества кластеризации и количества кластеров представлен в [3].

Глава 1

Методы и критерии оценки устойчивости кластеризации.

Понятие устойчивости кластеризации отличается от понятия устойчивости, введенном в математическом анализе. Здесь и в схожих работах используется статистическая устойчивость, основанная на применении симуляционных алгоритмов и процедур статистического моделирования.

В настоящем разделе описываются алгоритмы и критерии, которые используют искусственное возмущение для анализа поведения решения и его надежность. В кластерном анализе применение такого подхода не является новым, в работе [43] предлагается вносить искусственные возмущения в анализируемую выборку данных $X^{[n]}$ для анализа кластеризации. Способы возмущения данных предоставляются на выбор эксперта или исследователя. Уровень возмущений и их структура зависят от различного рода ошибок в данных, природы данных, качества, полноты и информативности выборки данных и других факторов. Цель использования случайных возмущений — моделирование потенциальных отклонений в выборке данных или проверка устойчивости/надежности решения и выводов при незначительных и естественных отклонениях во входных данных.

1.1. Процедура возмущения анализируемой выборки данных

Процесс возмущения исследуемой выборки данных $X^{[n]}$ требует статистических характеристик, прогнозных значений по изменению данных, оценки точности данных и других параметров в зависимости от источника получения и природы данных. Возмущение предлагается представлять в

виде случайной величины ξ, подчиняющейся вероятностному распределению с функцией распределения $F_\xi(z)$.

Получение и добавление возмущений к данным предлагается проводить следующим образом:

$$x_{ij}^{\sigma} = x_{ij} + \xi_j^i, \quad \forall x_i = (x_{i1}, \ldots, x_{im}) \in X^{[n]}, \quad \forall j = 1, \ldots, m, \tag{1.1}$$

где по каждой компоненте j используется своя функция распределения случайной величины $F_{\xi_j}(z)$, а индекс i в ξ_j^i указывает на то, что случайные величины добавляются независимо друг от друга.

Обозначим возмущенную по (1.1) выборку данных $X^{[n\sigma]}$. Количество элементов множества $X^{[n\sigma]} \subset X$ совпадает с $X^{[n]} \subset X$, но это не является строгим ограничением (к процедуре возмущения данных может быть добавлена дополнительная процедура выбора подмножества случайным образом, которая используется в индексах устойчивости кластеризации [43—45]). Функция распределения случайной величины определяется как функция $F_\xi(z) = F_{\xi_1}(z_1) \times \cdots \times F_{\xi_m}(z_m)$ и ξ является m-мерным вектором. Компоненты функции F_ξ могут представлять различные случайные распределения с различными параметрами, определяющими уровень возмущений по компоненте.

Функция F_ξ отражает неточности в данных или может представлять возможные изменения в данных (прогноз изменения в виде распределения вероятностей). Таким образом, сущность искусственного возмущения данных является оправданной и дополняет процедуру кластерного анализа.

Алгоритм возмущения исходной выборки данных не ограничивается процедурой прибавления шумового элемента. В зависимости от природы данных и знаний о характере ошибок или потенциальном изменении данных внесение шума может отличаться от указанного в настоящем разделе.

1.2. Имитационный алгоритм

Исходная выборка данных представлена в единственном экземпляре, процедура возмущения позволяет искусственно создать R наборов данных для исследования. Вспомогательные выборки используются в процессе имитаций задачи кластеризации на новых данных. На каждой итерации имитационного алгоритма рассчитываются матрицы кластеризации $\{C^{kr}\}_{k\in K}$ с элементами, определяемые формулой (1), для каждого рассматриваемого значения количества кластеров $k \in K$ и фиксированного на всех итерациях алгоритма кластеризации $\alpha(X^{[n]}, k)$ (или множество рассматриваемых алгоритмов кластеризации) на новом наборе данных, созданном на основе возмущенных данных $X^{[n\sigma_r]}$, где $r = 1, \ldots, R$ — номер итерации возмущения данных. Заметим, что возмущения на каждой итерации вносятся независимо от номера итерации, номер итерации используется с целью отделения одной итерации от другой.

Определение. Кластеризация C^k на выборке данных $X^{[n]}$ называется базовой, где $k \in K$.

Базовые кластеризации представляют результаты кластеризации на невозмущенных данных для каждого значения количества кластеров $k \in K$. Эти кластеризации используются в имитационном моделировании в качестве эталонной кластеризации для сравнения с соответствующими кластеризациями на возмущенных данных. Сравнение кластеризаций на каждой итерации r производится посредством заранее выбранной метрики (например, одной из рассмотренных функций из постановки задачи). Выбор метрики предлагается осуществлять в комбинации с уровнем внесенных возмущений. Если шум вносит большие изменения в структуре данных (величина шума сравнима с величиной данных), то целесообразно использовать непрерывную функцию оценки сходства кластеризаций. Если шум является незначительным по сравнению с объектами исследуемой выборки

данных, то грубая функция сравнения (например: (5)) будет достаточной для оценки близости кластеризаций.

В работах [13; 15] автором настоящей ВКР предлагаются два подхода проведения имитаций данных для кластеризаций.

Первый способ. Алгоритм проведения имитаций данных, в котором на каждой итерации возмущенное множество определяется как $X^{[n\sigma_r]}$, т.е. возмущенное множество, полученное на основе множества $X^{[n]}$ путем поэлементного добавления случайного возмущения (1.1). Предпосылками к оправданности таких операций над данными является предположение, что структура исходных данных должна совпадать или иметь незначительные расхождения со структурой «оправданно» возмущенных данных по метрике сравнения кластеризаций.

При многократном сравнении базовых кластеризаций с соответствующими по k кластеризациями на возмущенных данных получаем выборку расстояний (мер близости) между кластеризациями. На основе выборки расстояний между кластеризациями происходит статистическое принятие решения о статистической устойчивости кластеризации выборки $X^{[n]}$ на множестве количества кластеров K.

Второй способ. Алгоритм проведения имитаций основан на идее расширяющегося множества путем объединения множества $X^{[n]}$ с множествами $\{X^{[n\sigma_r]}\}_{r=1}^{R}$, т.е. на каждой итерации r множество для кластеризации представляется как $X^{[nr]} = X^{[n]} \cup X^{[n\sigma_1]} \cup \cdots \cup X^{[n\sigma_r]}$. В данном подходе возмущенные данные используются для «выжигания» неслучайных уплотнений данных, т.е. в результате расширения множества образуются плотные структуры (множества объектов), обозначающие наличие кластера. Принятие решений осуществляется на основе сравнения результатов кластеризаций на $X^{[n]}$ и $X^{[nr]}$ для соответствующих k.

В работе [43] предлагается процедура имитационного моделирования, основанная на сравнении результатов кластеризации на пересечениях слу-

чайных подмножеств множества $X^{[n]}$. Автором настоящей ВКР предлагается расширить процедуру моделирования вспомогательных выборок данных случайными возмущениями.

Алгоритмы имитации являются промежуточными стадиями в определении устойчивой кластеризации. Количество кластеров k в устойчивой кластеризации будем называть устойчивым. Определение устойчивого количества кластеров осуществляется на основе одного из статистических критериев определения устойчивой кластеризации (1.3) или (1.5).

Общий алгоритм имитационного моделирования состоит из трех основных этапов:

1. Внесение возмущений в элементы множества $X^{[n]}$;

2. Кластеризация;

3. Сравнение результатов кластеризации.

В результате имитационного моделирования получаем набор множеств значений $\{T_k\}_{k\in K}$ из R элементов. Элементом множества $t_{kr} \in T_k$ является мера сходства базовой кластеризации C^k и кластеризации C^{kr} (примеры функций сравнения кластеризаций: (2)—(5))

1.3. Bootstrapping множеств $\{T_k\}_{k\in K}$

Вычислительная сложность алгоритмов кластеризации в большинстве случаев зависит от количества объектов в исследуемой выборке, например, для k-средних вычислительная сложность линейна в зависимости от количества элементов $O(n)$, для k-ближайших соседей — $O(m^2)$, для алгоритма CURE — $O(m^2 logm)$, для алгоритма DBSCAN — (n^2). Обзор вычислительной сложности методов кластеризации представлен в работе [46]. Количество итераций в имитационном алгоритме R зависит от объемов анализируемых данных и ограничений вычислительных ресурсов. Так как R может быть невелико и множество элементов в T_k недостаточно для проведения

качественного статистического анализа, предлагается использовать статистическую процедуру Bootstrap [47] для расширения выборки качественных характеристик.

Множество T_k представляет набор значения меры сходства кластеризаций на возмущенных данных с базовой кластеризацией. В анализе устойчивости кластеризаций используется противоположное значение сходства — уровень сменяемости кластеризаций.

Определение. Уровень или частота сменяемости кластеризации для количества кластеров k: $\nu_k = \frac{1}{R}\sum_{r=1}^{R} t_{kr}$.

Метод bootstrap используется для генерации множеств $\{T_k^l\}_{l=1}^L$ для каждого $k \in K$, где L — количество генераций. Для каждого множества T_k^l рассчитывается ν_k^l по определению уровня сменяемости, множество уровней сменяемости $\{\nu_k^l\}_{l=1}^L$.

Введем следующие выборочные статистики: $N_k = \frac{\nu_k + \sum_{l=1}^{L} \nu_k^l}{L+1}$ — выборочный средний уровень сменяемости кластеризации с числом кластеров k и $S_k^2 = \frac{(\nu_k - N_k)^2 + \sum_{l=1}^{L}(\nu_k^l - N_k)^2}{L}$ — несмещенная выборочная дисперсия, которые будут использоваться в описании критериев устойчивой кластеризации (1.3) и (1.5).

Алгоритм получения промежуточных статистических показателей:

Инициализация: K, $\{T_k = \varnothing\}_{k \in K}$

Функции: $d(C^k, C^{kr})$

Для k из K выполнять

  Для r из $\{1, \ldots, R\}$ выполнять

    Генерация $X^{[n\sigma_r]}$

    $t_{kr} = d(C^k, C^{kr})$

    $T_k = T_k \cup \{t_{kr}\}$

    $\nu_k = \frac{1}{R}\sum_{t \in T_k} t$

  Для l из $\{1, \ldots, L\}$ выполнять

    Генерация множеств T_k^l случайным

выбором R элементов с повторением из T_k

$\nu_k^l = \frac{1}{R}\sum_{t\in T_k^l} t,$

где $d(C^k, C^{kr})$ — функция сравнения кластеризаций (например, одна из (2) — (5)). В результате работы алгоритма получим множество уровней сменяемости, на основе которых можно вычислить N_k и S_k^2. Блок-схема процедуры возмущения данных и bootstrap представлены на Рис. 1.1.

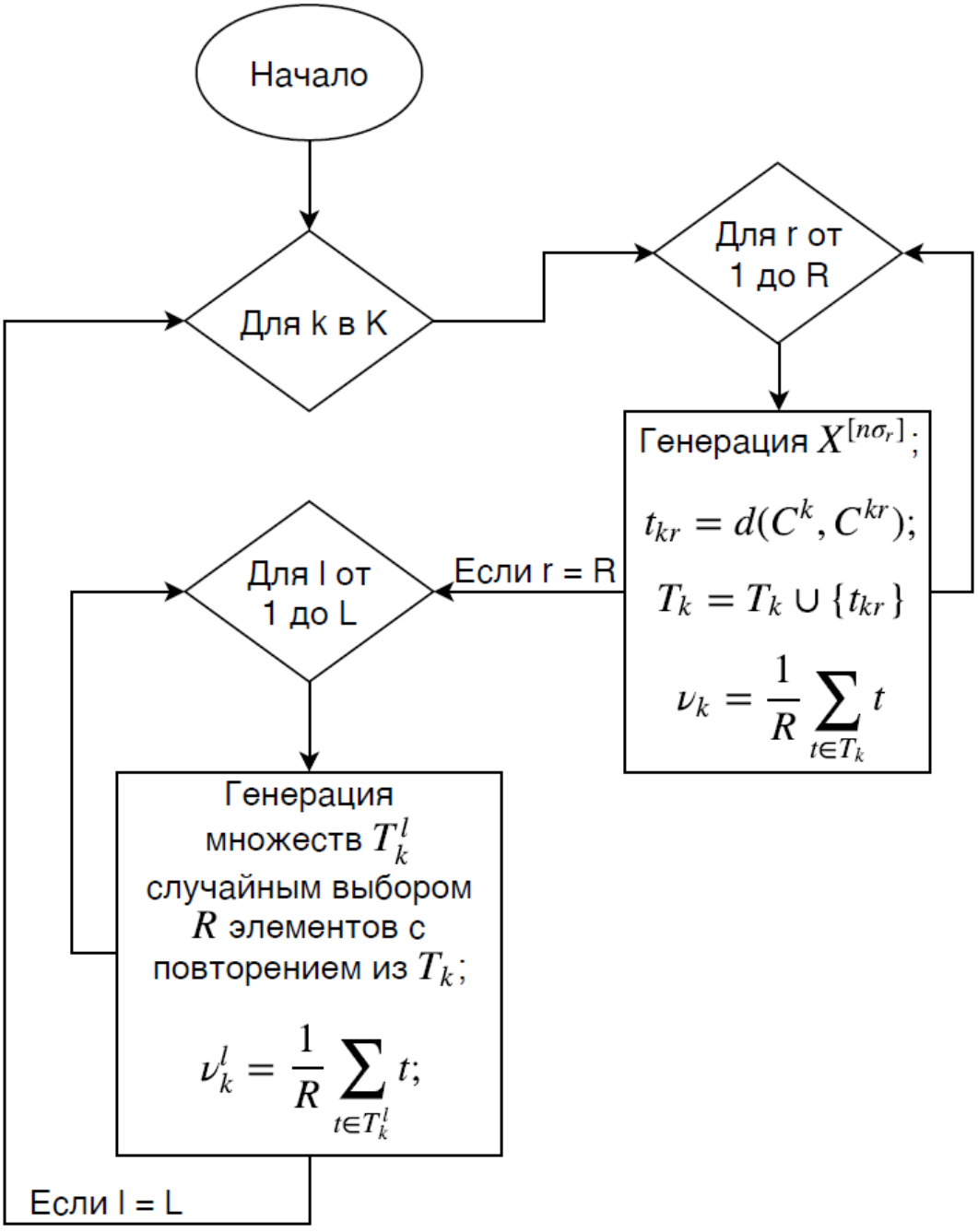

Рис. 1.1. Блок-схема возмущения и имитации данных

1.4. Выбор устойчивой кластеризации на основе теории рисков

В настоящем разделе рассматривается критерий выбора устойчивой кластеризации, основанный на теории рисков, а именно, с использованием Value at Risk (VaR) [48]. Стоимостная мера рисков VaR в применении к оценке портфеля активов — это величина убытков, которая не будет превышена с некоторой заранее заданной вероятностью (уровень доверия).

Не нарушая общности, предположим, что рассматриваются возрастающие метрики сравнения кластеризаций, т. е. чем больше значение принимает функция сравнения, тем больше сходство между кластеризациями.

Полученные в предыдущем разделе уровни сменяемости $\{\nu_k^l\}_{l=1}^L$ являются выборкой уровней сменяемости, которые получены из некоторой генеральной совокупности. Генеральная совокупность задается случайным распределением с замещенными истинными значений параметров на выборочные оценки, если первые не известны.

Рассмотрим случайную величину η — уровень сменяемости с функцией распределения $F_\eta(z)$. Тогда VaR случайной величины η для уровня доверия $\alpha \in (0,1)$ — это наибольшее число v, такое, что $P(\eta > v) = \alpha$, т. е. значение VaR для случайного уровня сменяемости η — это величина, которую η не превысит с вероятностью $1-\alpha$.

Применительно к оценке устойчивости кластеризации VaR используется в качестве индекса сменяемости кластеризации, который не будет превышен с уровнем доверия $1-\alpha$.

Моделирование функции распределения вероятностей $F_{\eta_k}(z)$ для каждого значения количества кластеров k предлагается проводить на основе множества $\{\nu_k^l\}_{l=1}^L$ и выборочных статистик N_k, S_k^2.

Пусть η_k — случайная величина уровня сменяемости ν_k с функцией распределения вероятностей $F_{\eta_k}(z)$ для количества кластеров k. Тогда фор-

мальное описание критерия выбора устойчивой кластеризации запишется в виде:

$$k_{stab} = arg \min_{k \in K}(sup\{z \in R : F_{\eta_k}(z) \leq 1 - \alpha\}),$$

т. е. наиболее устойчивой является та кластеризация, у которой величина уровня сменяемости, которая с вероятностью $(1 - \alpha)$ не будет превышена, является наименьшей на множестве K. Здесь R — обозначение множества вещественных чисел.

Рассмотрим пример для случая, когда η_k имеют нормальное распределение, тогда значение VaR находится из следующего выражения:

$$P(z > N_k + u_\alpha S_k) = \alpha, \tag{1.2}$$

где в качестве параметров нормального распределения используются замещенные параметры (статистические оценки, в случае отсутствия параметров генеральной совокупности), u_α — односторонняя α-квантиль стандартного нормального распределения. Рассмотрим следующие преобразования выражения (1.2):

$$P(z > N_k + u_\alpha S_k) = 1 - P(z \leq N_k + u_\alpha S_k) = \alpha,$$

$$P(z \leq N_k + u_\alpha S_k) = 1 - \alpha,$$

$$P(z \leq N_k - u_{1-\alpha} S_k) = 1 - \alpha.$$

С учетом нормального распределения частот сменяемости критерий выбора можно переформулировать следующим образом:

$$k_{stab} = arg \min_{k \in K}(N_k - u_{1-\alpha} S_k), \tag{1.3}$$

а значение VaR_k или индекс устойчивости определяется как

$$VaR_k = N_k - u_{1-\alpha} S_k. \tag{1.4}$$

Предложенный в настоящем разделе критерий выбора устойчивой кластеризации оценивает уровень сменяемости кластеризации с уровнем доверия $1-\alpha$ и предлагает выбирать кластеризацию с наименьшим VaR. Значение VaR в описанном критерии называется индексом устойчивости. Для оценки рисков изменения уровня сменяемости η_k возможно использовать другие оценки из теории рисков: CVaR [49], EVaR [50], DaR, CDaR.

1.5. Выбор устойчивой кластеризации на основе ожидаемого уровня сменяемости

Критерий, описанный в данном разделе, использует следующие предположения о статистической устойчивости кластеризации: количество кластеров k является устойчивым и определяет устойчивую кластеризацию C_k, если N_k и S_k^2 для этого количества кластеров являются наименьшими на $k \in K$. Задача состоит в минимизации двух параметров, т.е. задача многопараметрической оптимизации. Критерий был предложен автором ВКР в работе [11].

Пусть имеются две выборки $X = \{X_1, \ldots, X_n\}$ и $Y = \{Y_1, \ldots, Y_m\}$ из различных генеральных совокупностей с соответствующими плотностями распределений $f_\xi(x)$ и $f_\eta(x)$ и случайными величинами ξ и η. Пусть $\bar{x} > 0$ и $\bar{y} > 0$ — выборочные средние, S_1^2 и S_2^2 — несмещенные выборочные дисперсии для первой и второй выборок. Поставим задачу — сравнить две выборки по выборочному среднему и выборочной несмещенной дисперсии. Задача является многокритериальной, где лучшим решением будем считать выборку с наименьшим средним и дисперсией.

Введем в рассмотрение вероятности $P_1(|\xi| > \bar{x}+u) = 1-\int_{-\bar{x}-u}^{\bar{x}+u} \bar{f}_\xi(z)dz$ и $P_2(|\eta| > \bar{y}+u) = 1-\int_{-\bar{y}-u}^{\bar{y}+u} \bar{f}_\eta(z)dz$, где $u \geqslant 0$ — управляющий параметр, $\bar{f}_\xi(z)$ и $\bar{f}_\eta(z)$ — плотности распределений генеральных совокупностей с замещенными неизвестными истинными значениями параметров на статисти-

ческие оценки [51]: выборочными средними $\bar{x}$ и $\bar{y}$ и дисперсиями S_1^2 и S_2^2 соответственно.

Пусть $P_1(|\xi| > \bar{x} + u)$ и $P_2(|\eta| > \bar{y} + u)$ — вероятности отклонения значений более, чем на u. Для сравнения двух выборок предлагается критерий:

$$\min(\bar{x}P_1(|\xi| > \bar{x} + u), \bar{y}P_2(|\eta| > \bar{y} + u)).$$

Идея для критерия взята из работы [52], в которой описано именное правило — Littlewood's rule, применяемое в теории управления прибылью.

В случае оценки устойчивости кластеризации критерий сравнения выборок применяется к множеству $\{\nu_k^l\}_{l=1}^L$. Как было обозначено в Разделе 1.4, η_k — случайная величина уровня сменяемости для количества кластеров k с соответствующей плотностью распределения вероятностей $\bar{f}_{\eta_k}(x)$, т. е. с замещением истинных параметров оценочной выборочной средней N_k и оценочной дисперсией S_k^2. Кроме того, будем рассматривать $P_{\eta_k}(\eta_k \leq N_k + u)$, потому что отклонение уровня сменяемости кластеризаций в левую сторону является отклонением в сторону повышения устойчивости. Вероятность отклонения частоты сменяемости больше, чем на u, запишется как

$$P_{\eta_k}(\eta_k > N_k + u) = 1 - P_{\eta_k}(\eta_k \leq N_k + u) = 1 - \int_{-\infty}^{N_k+u} \bar{f}_{\eta_k}(x)dx.$$

Таким образом, критерий выбора устойчивой кластеризации и устойчивого количества кластеров примет вид:

$$k_{stab} = arg\min_{k \in K}[N_k - N_k P_{\eta_k}(\eta_k \leq N_k + u)], \tag{1.5}$$

а индекс устойчивости кластеризации для количества кластеров равным k будет иметь вид:

$$\sigma_k = N_k - N_k P_{\eta_k}(\eta_k \leq N_k + u) \tag{1.6}$$

Рассмотрим частный случай, когда $\{\nu_k^l\}_{l=1}^L$ — выборки из нормального распределения с замещенными параметрами распределения на статистические оценки и $\{\bar{f}_{\eta_k}(x)\}_{k\in K}$ — плотности распределения.

Тогда критерий с учетом обозначений запишется следующим образом:

$$arg\min_{k\in K} N_k P_{\eta_k}(x > N_k + u)$$

или

$$arg\min_{k\in K} N_k - N_k \int\limits_{-\infty}^{N_k+u} \frac{1}{S_k\sqrt{2\pi}} e^{-\frac{(z-N_k)^2}{2S_k^2}} dz.$$

Предложенный критерий — это ожидание отклонения уровня сменяемости ν_k больше, чем на u, в сравнении с выборочным средним N_k. Значение параметра u является управляющей функцией.

Выбор управляющего параметр u является задачей эксперта, аналогично уровню доверия α в критерии основанном на VaR. Это может быть среднее значение среднеквадратичных отклонений $u_{avg} = \sum_{k\in K} \frac{S_k}{|K|}$ (минимальное, максимальное и другие функции от S_k^2), что позволяет учесть разброс по уровням сменяемости в выборках частот сменяемости для каждого количества кластеров $k \in K$.

Глава 2

Численный эксперимент и сравнение с существующими результатами

Настоящая глава содержит результаты применения предложенных в Главе 1 индексов качества кластеризации, основанных на оценке устойчивости кластеризации (1.3) и (1.5). Для этого предлагается рассмотреть три типа наборов данных: первые два типа наборов данных являются синтетическими (созданные с использованием известных распределений случайных величин или определенной геометрической структуры), где один из типов состоит из Гауссовых кластеров (выборка данных, полученная из нескольких Гауссовых распределений с разными параметрами, Рис. 2.1), а другой тип данных состоит из кластеров определенной формы и размера в двумерном пространстве (Рис. 2.2). Третья группа данных состоит из реальных выборок, размещенных в репозитории машинного обучения UCI [53].

На Рис. 2.1, первая выборка данных состоит из 5 подвыборок двумерного нормальных распределений с незначительным пересечением между собой (5Gauss). Второй набор данных сформирован из 4 гауссианов с различной степенью нахлеста (4Gauss), третье множество состоит из 6 гауссианов (6Gauss), где 2 из них полностью накладываются друг на друга с одинаковыми средними, но разными ковариационными матрицами.

Следующие два набора данных имеют кластеры различной формы, плотности и размера (Рис. 2.2). Первый набор данных состоит из двух концентрических окружностей (2CRings), а второй набор данных сформирован из двух полу-колец (2HRings).

Три набора данных взяты из репозитория машинного обучения UCI [53]: Wisconsin Diagnostic Breast Cancer (Wdbc), Iris и Wine.

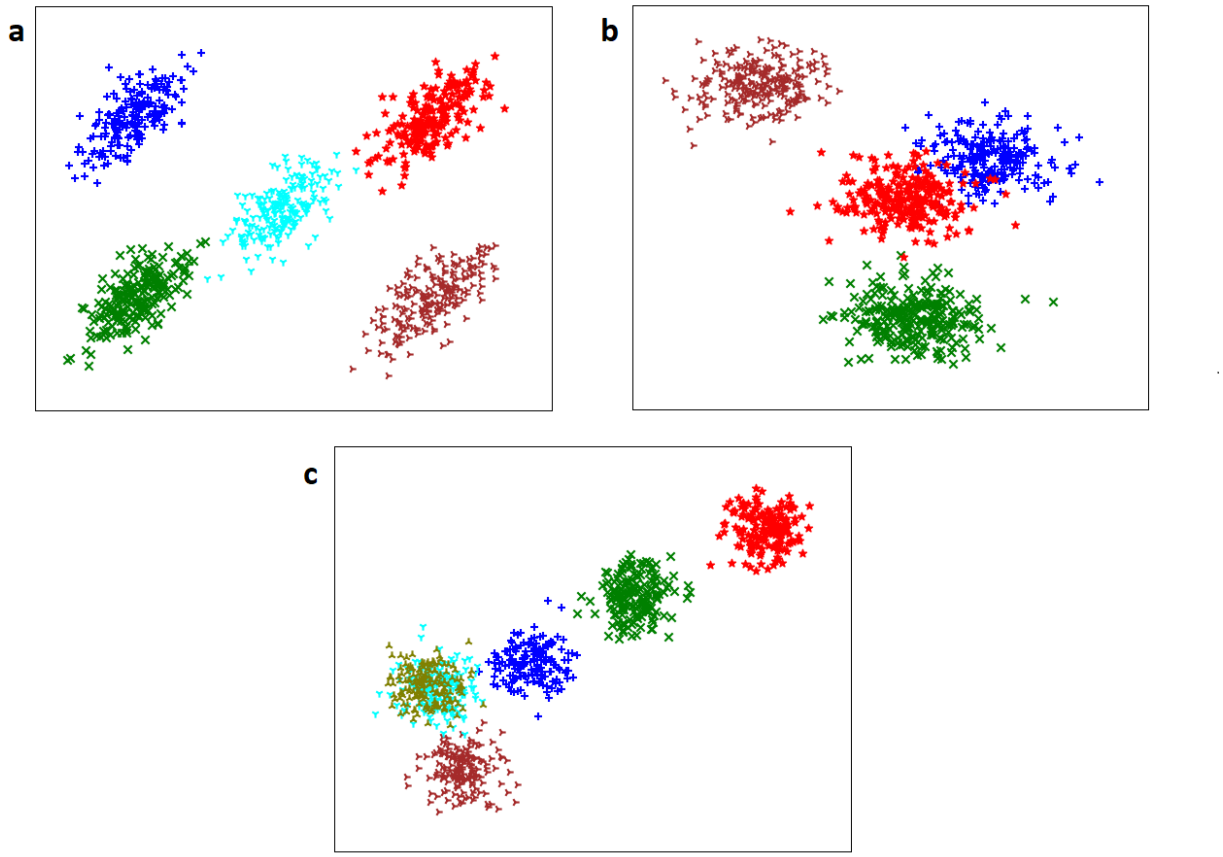

Рис. 2.1. Искусственные данные первого типа. Слева на право: a) 5 Gauss b) 4 Gauss и c) 6 Gauss

Набор данных Wdbc содержит 569 объектов с 30 признаками, полученными оцифровыванием изображений, сделанных при помощи тонкой аспираторной иглы (FNA). Набор данных разделен на 2 класса.

В наборе данных Iris («Ирисы Фишера») содержатся три группы объектов по 50 экземпляров, где каждая группа представляет собой разновидность ирисов (iris virginica, iris setosa и iris versicolor). У каждого экземпляра имеются четыре характеристики (признака): длина и ширина наружной доли околоцветника, длина и ширина внутренней доли околоцветника. Одна из групп линейно отделима от двух остальных, но последние два не отделимы друг от друга (имеют пересечение).

Набор данных Wine. Этот набор данных является результатом химического анализа вин, выращенных в одном регионе в Италии, но произ-

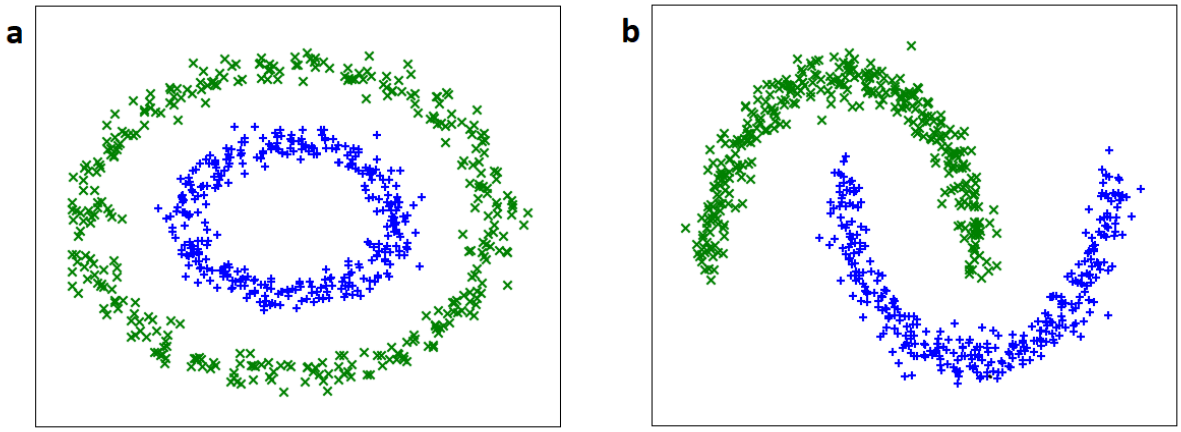

Рис. 2.2. Искусственные данные второго типа. Слева на право: a) 2 CRings b) 2 HRings

веденных на основе трех различных сортов винограда. Каждый объект в выборке имеет 13 признаков, всего 178 объектов в выборке. В эксперименте рассматриваются три различных алгоритма кластеризации: k-средних (англ. k-means), EM-алгоритм, основанный на максимизации функции правдоподобия, и иерархический алгоритм кластеризации. Эти алгоритмы соответствуют трем различным моделям кластеризации, где k-средних ищет компактные кластеры около среднего, EM-алгоритм ищет кластеры гауссовой формы, а иерархический алгоритм, основанный на оценках плотности данных и стратегии одиночной связи, подходит для кластеров любой структуры.

Исходный код программы и расширенные результаты эксперимента: для 4 функций сравнения кластеризаций, двух стратегий возмущения данных, четырех различных значений параметров u, — доступны на github.com[1].

2.1. Расчет уровней сменяемости кластеризаций

В экспериментах множество значений количества кластеров принятых к рассмотрению: $K = \{2, \ldots, 10\}$, где для каждого алгоритма кластеризации на каждом наборе данных производится расчет кластеризации и вычис-

[1] https://github.com/Hippy92/stable_clustering_estimation.git

ление уровня устойчивости кластеризации согласно (1.4) и (1.6) обозначенные VaR и σ_0 соответственно, управляющим параметром $u = 0$, функцией сравнения кластеризаций (2) [2], предполагается нормальное распределение выборок $\{\nu_k^l\}_{l=1}^L$ для $k \in K$. В процедурах оценки устойчивости кластеризации использован алгоритм возмущения данных, описанный в Разделе 1.1, количество генераций возмущенных данных $R = 100$, возмущения сгенерированы из непрерывного равномерного распределения с симметричным интервалом, где правая и левая границы (+) соответствуют доле от среднего значения в выборке данных по компоненте, рассматриваются случи с долей равной 1%. Количество итераций алгоритма bootstrap из Раздела 1.3 $L = 1000$, где генерация последовательностей $\{T_k^l\}_{l=1}^L$ осуществляется на основе равновероятного выбора элементов с повторением из T_k. В случаях, когда уровни сменяемости кластеризаций совпадают и являются минимальными, в рассмотрение берется кластеризация с наибольшим количеством кластеров.

В эксперименте рассматривались два индекса устойчивости (1.4) и (1.6), обозначенные VaR и σ_0 соответственно. Жирным шрифтом выделены оптимальные (устойчивые по соответствующему критерию) значения кластеров.

2.2. Синтетические данные

В Таблице 2.1 представлены результаты расчета уровней сменяемости на искусственных данных с использованием алгоритма кластеризации k-средних для индексов устойчивости (1.4) и (1.6). Заметим, что критерии (1.3) и (1.5) определили устойчивую кластеризацию с количеством кластеров, соответствующим истинному количеству кластеров для 5Gauss,

[2] Максимальное значение, которое может принимать функция это 1, в случае полного совпадения кластеризаций. Для применения данной функции в эксперименте введена модификация функции: 1 «минус» функция (2).

2HRing. В свою очередь, 6Gauss содержит 2 пересекающихся кластера, как следствие, алгоритм кластеризации не способен распознать их, поэтому устойчивая кластеризация 6Gauss с количеством кластеров равным 5 является верной (см. Рис. 2.1). Алгоритм кластеризации k-средних выбирает центр кластера и связывает с ним ближайшие объекты, поэтому структура кластеров 2CRings (см. Рис. 2.2) не может быть распознана этим алгоритмом кластеризации. Набор данных 4Gauss имеет наиболее устойчивые кластеризации с количеством кластеров 2 и 3, но уровень сменяемости кластеризации для $k = 4$ на 2 порядка ниже в сравнении с остальными (за исключением $k = 2$ и $k = 3$) и является следующей по величине значения индекса устойчивости кластеризацией после $k = 2$ и $k = 3$. Наличие пересечений между двумя кластерами в 4Gauss (см. Рис. 2.1) привело к колебаниям кластеризаций при многократном возмущении исходной выборки данных 4Gauss.

k	**5Gauss**		**4Gauss**		**6Gauss**		**2CRings**		**2HRings**	
	VaR	σ_0	VaR	σ_0	VaR	σ_0	VaR	σ_0	VaR	σ_0
2	0,0075	0,0034	**0**	**0**	**0**	**0**	0,1117	0,0463	**0,0021**	**0,0009**
3	0,0246	0,0071	**0**	**0**	0,0041	0,0015	**0,0351**	**0,0121**	0,0830	0,0349
4	0,0107	0,0040	0,0016	0,0007	0,0196	0,0051	0,0998	0,0387	0,0246	0,0092
5	**0,0023**	**0,0010**	0,1347	0,0620	**0**	**0**	0,0356	0,0163	0,0179	0,0054
6	0,0180	0,0068	0,1965	0,0949	0,0157	0,0066	0,0713	0,0303	0,0218	0,0096
7	0,0793	0,0337	0,2743	0,1327	0,0457	0,0190	0,3560	0,1706	0,0849	0,0387
8	0,0754	0,0323	0,2520	0,1196	0,1379	0,0646	0,4493	0,2194	0,0293	0,0127
9	0,0894	0,0392	0,2240	0,1078	0,2020	0,0940	0,4279	0,2048	0,0582	0,0262
10	0,0797	0,0371	0,1828	0,0849	0,2904	0,1399	0,3436	0,1648	0,0483	0,0218

Таблица 2.1. Индексы устойчивости для синтетических наборов данных полученные посредством использования алгоритма кластеризации k-средних

В Таблице 2.2 представлены значения индексов устойчивости кластеризаций на разных синтетических данных, полученные в результате применения EM-алгоритма. Из Таблицы 2.2 видно, что устойчивое количество кластеров соответствует правильному для 5Gauss, 2CRings, 2HRings. Ана-

логично результатам для k-средних на 6Gauss и 4Gauss, EM-алгоритм также получил устойчивую кластеризацию с $k = 5$ и $k = 3$ соответственно, при этом в последнем случае уровень сменяемости кластеров для $k = 4$ на 2 порядка ниже в сравнении с другими.

k	**5Gauss**		**4Gauss**		**6Gauss**		**2CRings**		**2HRings**	
	VaR	σ_0	VaR	σ_0	VaR	σ_0	VaR	σ_0	VaR	σ_0
2	0,0053	0,0024	**0**	**0**	**0**	**0**	**0,0140**	**0,0064**	**0,0005**	**0,0002**
3	0,0459	0,0149	**0**	**0**	0,0008	0,0003	0,0282	0,0108	0,0175	0,0077
4	0,0118	0,0033	0,0008	0,0003	0,0091	0,0024	0,0273	0,0092	0,2205	0,1022
5	**0,0016**	**0,0007**	0,0820	0,0387	**0**	**0**	0,0944	0,0447	0,0044	0,0020
6	0,0106	0,0042	0,1161	0,0540	0,0046	0,0019	0,0644	0,0270	0,0026	0,0011
7	0,0629	0,0266	0,1304	0,0615	0,0185	0,0073	0,2874	0,1393	0,0618	0,0285
8	0,0428	0,0185	0,1134	0,0541	0,0414	0,0176	0,3145	0,1463	0,0524	0,0232
9	0,0364	0,0153	0,0659	0,0308	0,0486	0,0219	0,3570	0,1669	0,0425	0,0192
10	0,0628	0,0293	0,0880	0,0410	0,0973	0,0440	0,2737	0,1282	0,0102	0,0048

Таблица 2.2. Индексы устойчивости для синтетических наборов данных полученные посредством использования EM-алгоритма кластеризации

Иерархический алгоритм кластеризации некорректно отработал на данных полученных из нормальных распределений (5Gauss, 4Gauss, 6Gauss, см. Рис. 2.1), и количество кластеров в устойчивых кластеризациях по предлагаемым критериям не соответствует истинным (см. Таб. 2.3). Но на данных негауссовой структуры 2CRings и 2HRings (см. Рис. 2.2) значение k в устойчивых кластеризациях совпадает с истинными.

2.3. Реальные данные

В Таблице 2.4 и Таблице 2.5 представлены результаты эксперимента на реальных данных (Wdbc, Iris, Wine) для каждого из трех рассматриваемых алгоритмов и двух индексов устойчивости (1.4) и (1.6). Устойчивые кластеризации для всех алгоритмов на данных Wdbc и Iris имеют одинаковое количество кластеров $k = 2$. Отметим, что данные Iris содержат два

k	**5Gauss**		**4Gauss**		**6Gauss**		**2CRings**		**2HRings**	
	VaR	σ_0	VaR	σ_0	VaR	σ_0	VaR	σ_0	VaR	σ_0
2	0,0302	0,0087	**0,0010**	**0,0004**	0,0575	0,0209	**0,0140**	**0,0039**	**0,0005**	**0,0797**
3	0,0131	0,0037	0,0022	0,0008	0,0482	0,0170	0,0282	0,0094	0,0175	0,1083
4	**0,0064**	**0,0026**	0,0585	0,0277	**0,0047**	**0,0018**	0,0273	0,0179	0,2205	0,1302
5	0,0070	0,0029	0,1310	0,0633	0,0085	0,0034	0,0944	0,0312	0,0044	0,0974
6	0,0365	0,0154	0,1877	0,0897	0,0988	0,0469	0,0644	0,0315	0,0026	0,0985
7	0,0953	0,0429	0,2599	0,1249	0,1492	0,0716	0,2874	0,0475	0,0618	0,1082
8	0,1383	0,0639	0,3225	0,1568	0,1646	0,0791	0,3145	0,0657	0,0524	0,0809
9	0,1334	0,0613	0,3358	0,1631	0,2132	0,1019	0,3570	0,0643	0,0425	0,1114
10	0,1971	0,0928	0,3495	0,1705	0,2702	0,1303	0,2737	0,0383	0,0102	0,1054

Таблица 2.3. Индексы устойчивости для синтетических наборов данных полученные посредством использования иерархического алгоритма кластеризации

пересекающихся друг с другом кластера, но значения Var и σ_0 для $k = 3$ и алгоритмов кластеризации k-средних и иерархической кластеризации являются следующими по величине после $k = 2$. С точки зрения устойчивости (несменяемости кластеризаций), результаты являются корректными, и в случае с Iris кластеризация с $k = 2$ является более устойчивой, чем при $k = 3$.

k	**k-средних**						**EM-алгоритм**					
	Wdbc		**Iris**		**Wine**		**Wdbc**		**Iris**		**Wine**	
	VaR	σ_0	VaR	σ_0	VaR	σ_0	VaR	σ_0	VaR	σ_0	VaR	σ_0
2	**0,001**	**0**	**0**	**0**	**0,004**	**0,002**	**0**	**0**	**0**	**0**	0,037	0,014
3	0,012	0,006	0,007	0,003	0,008	0,004	0,002	0,001	0,035	0,016	0,018	0,008
4	0,013	0,005	0,047	0,020	0,100	0,044	0,012	0,006	0,006	0,002	**0,012**	**0,006**
5	0,031	0,014	0,019	0,008	0,034	0,012	0,018	0,008	0,004	0,002	0,029	0,012
6	0,085	0,033	0,047	0,022	0,024	0,010	0,044	0,018	0,035	0,014	0,084	0,037
7	0,252	0,117	0,076	0,034	0,041	0,018	0,105	0,043	0,034	0,015	0,188	0,084
8	0,118	0,051	0,076	0,034	0,083	0,035	0,182	0,077	0,087	0,041	0,161	0,073
9	0,100	0,042	0,198	0,095	0,205	0,091	0,033	0,015	0,168	0,080	0,215	0,097
10	0,135	0,056	0,183	0,086	0,106	0,046	0,111	0,048	0,180	0,085	0,124	0,056

Таблица 2.4. Индексы устойчивости для наборов данных Wdbc, Iris и Wine полученные посредством использования алгоритмов кластеризации k-средних и EM-алгоритма

k	**Wdbc**		**Iris**		**Wine**	
	VaR	σ_0	VaR	σ_0	VaR	σ_0
2	**0,088**	**0,040**	**0**	**0**	**0,086**	**0,036**
3	0,276	0,129	0,021	0,010	0,235	0,107
4	0,215	0,099	0,032	0,015	0,182	0,083
5	0,184	0,084	0,036	0,017	0,270	0,126
6	0,226	0,103	0,090	0,041	0,251	0,119
7	0,325	0,154	0,144	0,068	0,248	0,117
8	0,346	0,166	0,135	0,063	0,270	0,128
9	0,335	0,159	0,181	0,086	0,247	0,116
10	0,360	0,170	0,217	0,104	0,216	0,101

Таблица 2.5. Индексы устойчивости для наборов данных Wdbc, Iris и Wine полученные посредством использования иерархический алгоритм кластеризации

Устойчивые кластеризации набора данных Wine не совпали с истинными (Таблица 2.1, Таблица 2.5), но для алгоритмов кластеризации k-средних и EM-алгоритма значение уровня устойчивости для кластеризации с $k = 3$ является следующими после наиболее устойчивой кластеризации. Заметим, что наиболее устойчивые кластеризации для k-средних, иерархической кластеризации и EM-алгоритма отличаются.

2.4. Сравнение результатов с другими индексами устойчивости кластеризации

В настоящем разделе представлены результаты сравнения предлагаемого подхода к определению числа кластеров на основе устойчивости с тремя другими индексами валидации количества кластеров, которые так же основаны на устойчивости. Алгоритмы сравниваются на всех наборах данных, введенных в рассмотрение в начале Раздела 2.

Первый алгоритм [20] использует то же определение матрицы кластеризации C и ее свойства, описанные в Постановке задачи. В процедурах алгоритма из анализируемой выборки выбираются случайным образом две

пересекающиеся подвыборки с количеством элементов в долевом соотношении к исходной выборке $f \geq 0.6$. Данная процедура повторяется N раз. На каждой итерации происходит сравнение кластеризаций на пересечении подвыборок. Основное предположение данного алгоритма оценки количества кластеров — сохранение структуры данных на случайных подмножествах. Метрики сравнения кластеризаций описаны в Постановке задачи. В Таблицах 2.6—2.8 результаты оптимального количества кластеров по критерию [20] обозначены Ben-Hur.

Второй алгоритм, который используется для сравнения результатов, представлен в [44]. Этот алгоритм делит выборку случайным образом на две непересекающиеся подвыборки N раз, в дальнейшем одна из подвыборок используется в качестве обучающей выборки, а вторая — для проверки. Используя метод перекрестной валидации, обе подвыборки разбиваются на k кластеров одним и тем же алгоритмом клстеризации. В дальнейшем оценивается, насколько хорошо обучающая выборка определяет принадлежность к кластерам объектов тестовой выборки. В тестовой выборке данных фиксируются объекты, лежащие в одном кластере (полученные прямой кластеризацией), объекты которого сравниваются с метками классов (кластеров) этих же объектов, полученных в результате прогнозирования на основе обучающей выборки. В Таблицах 2.6—2.8 результаты оптимального количества кластеров по критерию [44] обозначены Tibshi.

Третий алгоритм, включенный в сравнительный анализ представлен в работе [45]. Процедура и концепция алгоритма схожа с алгоритмом Tibshi, отличие заключается в методе сравнения кластеров, полученных в процессе обучения и при помощи методов кластеризации. В Таблицах 2.6—2.8 результаты оптимального количества кластеров по критерию [45] обозначены Pascual.

Из Таблиц 2.6—2.8 для 5Gauss оптимальное количество кластеров для алгоритмов кластеризации k-средних и EM-алгоритма совпадают, но для

Индекс	**5Gauss**	**4Gauss**	**6Gauss**	**2CRings**	**2HRings**	**Wdbc**	**Iris**	**Wine**
VaR/σ_0	5	2/3	2/5	3	2	2	2	2
Ben-Hur	5	3	3/5	4	2	2	2	2
Tibshi	5	3	5	4	2	2	2	2
Pascual	5	3	5	3	2	2	2	4

Таблица 2.6. Устойчивые значения количества кластеров для алгоритма кластеризации k-средних

случая с иерархическим алгоритмом VaR/σ_0 и Pascual возникают неточности с определением кластеров гауссовой структуры.

В наборе данных 4Gauss только в одном случае было определено правильное количество кластеров в случае применения EM-алгоритма кластеризации и по критерию Pascual. В остальных случаях оптимальным было $k = 2$ или $k = 3$, можно предположить, что пересекающиеся два кластера в 4Gauss неразделимы на уровне машинного распознавания.

Индекс	**5Gauss**	**4Gauss**	**6Gauss**	**2CRings**	**2HRings**	**Wdbc**	**Iris**	**Wine**
VaR/σ_0	5	2/3	2/5	2	2	2	2	4
Ben-Hur	5	3	5	5	4	2	2	2
Tibshi	5	2	5	2	4	2	2	2
Pascual	5	4	5	2	4	2	2	4

Таблица 2.7. Устойчивые значения количества кластеров для EM-алгоритма кластеризации

Результаты на выборке данных 6Gauss зависят от алгоритма кластеризации, но в основном оптимальное количество кластеров является истинным, т. к. два кластера в 6Gauss неразделимы.

Выборки данных 2CRings и 2HRings имеют негауссову структуру, которую алгоритмы кластеризации k-средних и EM-алгоритм не адаптированы распознавать, поэтому получение хороших результатов по данным алгоритмам кластеризации является случайностью. Алгоритм иерархической кластеризации, напротив, способен «хорошо» распознать такие формы кластеров как в 2CRings и 2HRings, что видно в Таблице 2.8.

Оптимальные значения количества кластеров на выборках реальных данных совпадают между разными критериями за исключением Pascual и VaR/σ_0 на выборке Wine. Заметим, что отклонения происходят из-за алгоритма кластеризации, т.е. алгоритм кластеризации является параметром, который влияет на итоговый результат.

Индекс	**5Gauss**	**4Gauss**	**6Gauss**	**2CRings**	**2HRings**	**Wdbc**	**Iris**	**Wine**
VaR/σ_0	4	2	4	2	2	2	2	2
Ben-Hur	5	3	5	2	2	2	2	2
Tibshi	5	3	5	2	2	2	2	2
Pascual	3	3	3	2	2	2	2	2

Таблица 2.8. Устойчивые значения количества кластеров для иерархического алгоритма кластеризации

В Таблицах 2.6—2.8 можно видеть, что VaR/σ_0 определяет правильное количество кластеров лучше на данных негауссовой структуры типа 2CRings и 2HRings, чем остальные рассмотренные алгоритмы определения числа кластеров независимо от алгоритма кластеризации. Результаты эксперимента на 4Gauss, 5Gauss, 6Gauss и реальных данных не выявили преимуществ VaR/σ_0 перед рассмотренными критериями.

Глава 3

Приложение критериев устойчивости кластеризации к задаче о размещении хабов в сети

Исследования сетей имеет непосредственное влияние на такие отрасли индустрии, как перевозка пассажиров и грузов наземным/морским/авиа транспортом, почтовые доставки, телекоммуникационное обслуживание и др. Такие сети часто содержат большое количество пар отправитель—получатель для обслуживания, где прямые соединения между узлами сети не всегда возможны ввиду географических, экономических или технических ограничений. Введение сети хабов (англ. hub) призвано значительно сократить количество связей в сети и уменьшить размерность задачи через консолидацию, перегруз или распределение потоков в сети. Сокращение затрат достигается в результате маршрутизации потоков сети через один или более хабов. Задача размещения хабов состоит в назначении набора узлов сети хабами и построении связей между направлениями отправитель–получатель и хабами оптимальным образом.

Основой теории размещения хабов (англ. Hub Location Problem (HLP)) является работа О'Келли [54]. Первым этапом развития HLP было формулирование и классификация задач: ограниченная/неограниченная пропускная способность хабов, характер связи между направлениями и хабами, дискретная/вещественная постановка задачи, размещение одного/p/не фиксированного количества хабов в сети и др. Широкий обзор постановок задач HLP и подходов к их решению представлен в работе [55].

Одним из поздних направлений в HLP является постановка робастной (устойчивой) сети хабов. Так как задача HLP решается на уровне страте-

гического планирования, на длительный период времени, и принятие решений происходит в условиях неопределенности (спрос, стоимость открытия хаба, транспортные издержки и др.), то возникает потребность в нахождении устойчивого решения.

В работе [56] предлагается деление робастных концепций HLP по виду представления неопределенности: заданный набором сценариев с указанием вероятности появления сценария, неопределенность, выраженная в виде ограничений (в виде интервалов или выпуклого многогранника). Ко второму виду относятся постановки, описанные в работах [56; 57]. Концепция робастности, представленная в настоящем разделе, основана на рассмотрении набора сценариев, как и в работах [16; 58], т. е. первый вид робастных HLP.

Основные результаты нахождения устойчивого размещения хабов в сети представлены в работах автора в соавторстве с научным руководителем [16; 17].

3.1. Описание задачи UMApHLP

В работе для демонстрации предлагаемой робастной постановки HLP используется формулировка задачи из [59] с фиксированным количеством хабов p (англ. Uncapacitated Multiple Allocation p-Hub Location Problem (UMApHLP)), но результаты настоящей работы не ограничиваются указанной моделью и могут быть распространены на другие постановки HLP.

В постановке задачи используются стандартные предположения: сеть хабов является полным графом, нет прямых соединений отправитель—получатель, пропускная способность хабов не ограничена, перегруз разрешен только на хабах.

В формулировании задачи используются следующие обозначения: $N = \{1, \ldots, n\}$ — набор узлов сети, множество потенциальных хабов $K \subset N$,

расстояние между $i, j \in N$ обозначено как d_{ij}, a_k — стоимость установки хаба в узле сети $k \in K$; стоимости перегруза, консолидации и распределения еденици потока на единицу расстояния обозначены как α, χ и δ соответственно; w_{ij} — поток, направленный от отправителя $i \in N$ к получателю $j \in N$ (спрос на направление). Так как потоки должны проходить как минимум через один хаб, а использование более двух хабов является избыточным, то стоимость перемещения единицы потока по маршруту представляется как $c_{ijkm} = \chi d_{ik} + \alpha d_{km} + \delta d_{mj}$, где $i, j \in N$ — отправитель–получатель, а $k, m \in K$ — хабы в маршруте, p — кол-во хабов, которое необходимо выбрать.

Математическая постановка базовой задачи HLP следующая:

$$\min \quad \sum_{k \in K} a_k y_k + \sum_{i \in N} \sum_{j \in N} \sum_{k \in K} \sum_{m \in K} c_{ijkm} x_{ijkm} \tag{3.1}$$

при ограничениях

$$\sum_{m \in K, m \neq k} x_{ijmk} + \sum_{m \in K} x_{ijkm} \leqslant w_{ij} y_k, \quad i, j \in N, k \in K, \tag{3.2}$$

$$\sum_{k \in K} \sum_{m \in K} x_{ijkm} = w_{ij}, \quad i, j \in N, \tag{3.3}$$

$$\sum_{k \in K} y_k = p, \tag{3.4}$$

$$x_{ijkm} \geqslant 0, \quad i, j \in N, k, m \in K, \tag{3.5}$$

$$y_k \in \{0, 1\}, \quad k \in K, \tag{3.6}$$

где y_k — бинарная переменная, принимающая значение 1, если $k \in K$ выбран в качестве хаба, 0 — в противном случае, а вещественная переменная x_{ijkm} соответствует потоку из $i \in N$ в $j \in N$ через хабы $k, m \in K$.

Целевая функция (3.1) представляет собой общие затраты сети для минимизации. Неравенства (3.2) обеспечивают маршрутизацию потоков только через выбранные хабы, а ограничения (3.3) гарантируют, что весь исходящий поток будет доставлен получателю, уравнение (3.4) обеспечивает выполнение выбора p хабов.

Постановка задачи (3.1) — (3.6) используется как основа для описания процедуры оценки устойчивости сети хабов к неопределенности в спросе в задаче HLP. Обозначим (3.1) — (3.6) как функцию $G(W, p)$ зависящую от матрицы спроса W и количества хабов p, где значения функции — это вектор оптимальных хабов к размещению $G(W, p) = (y_1, \ldots, y_{|K|})$, а остальные параметры модели (3.1) — (3.6) фиксированы.

3.2. Сравнение двух результатов решения задачи UMApHLP

Процедура нахождения устойчивого количества хабов совпадает с алгоритмом нахождения устойчивой кластеризации в Главе 1. Обозначим возмущенную матрицу спроса W^{σ_r}, аналогично $X^{[n\sigma_r]}$ в Разделе 1.2, где $r \in \{1, \ldots, R\}$.

Для сравнения результатов значений функций $G(W, p)$ и $G(W^{\sigma_r}, p)$ можно использовать функции сравнения кластеризаций из Постановки задачи путем представления $G(W, p)$ и $G(W^{\sigma_r}, p)$ в виде результата кластеризации на 2 кластера, где в одном из кластеров будет p хабов (выбранные хабы к размещению), а во втором $(|K| - p)$ хабов.

Введем следующую функцию сравнения близости значений двух функций $G(W, p)$ и $G(W^{\sigma_r}, p)$, т. е. сравнение двух результатов решения задач (3.1) — (3.6) с одинаковым количеством хабов p, но разными матрицами

спроса:

$$d(G(W,p),G(W^{\sigma_r},p)) = \sum_{k\in K}(y_k \neq y_k^{\sigma_r}) \tag{3.7}$$

Функция (3.7) может принимать только целочисленные значения.

Оценка устойчивости количества хабов p основана на многократном решении задачи $G(W,p)$ с возмущением спроса W. Рассмотрение возмущений спроса связано с неопределенностью, возникающей при планировании сети хабов. В качестве источников неопределенности в задаче HLP могут быть рассмотрены также стоимость открытия хаба a_k, стоимость перемещения единицы потока c_{ijkm} или расстояние между узлами сети d_{ij}.

Алгоритм нахождения устойчивого $p \in \{p_{\min},\dots,p_{\max}\}$, основанный на процедуре, описанной в Разделе 1, и критерии оптимальности VaR (3.8) имеет следующий вид:

Инициализация: W, $p_{\min}$, $p_{\max}$, $\{T_p = \varnothing\}_{p=p_{\min}}^{p_{\max}}$
Функции: $G(W,p)$, $d(G(W,p),G(W^{\sigma_r},p))$
Для p из $\{p_{\min},\dots,p_{\max}\}$ выполнять
  Для r из $\{1,\dots,R\}$ выполнять
    Генерация W^{σ_r}
    $t_{pr} = d(G(W,p),G(W^{\sigma_r},p))$
    $T_p = T_p \cup \{t_{pr}\}$
    $\nu_p = \frac{1}{R}\sum_{t\in T_p} t$
  Для l из $\{1,\dots,L\}$ выполнять
    Генерация множеств T_p^l случайным
    выбором R элементов с повторением из T_p
    $\nu_p^l = \frac{1}{R}\sum_{t\in T_p^l} t$

В алгоритме L обозначает количество итераций bootstrap из Раздела 1.3. В результате работы алгоритма получим частоты сменяемости хабов

ν_p^l. Статистические характеристики сменяемости для $\forall p \in \{p_{\min}, \ldots, p_{\max}\}$ вычисляются по формулам из Раздела 1.3: $N_p = \frac{\nu_p + \sum_{l=1}^{L} \nu_p^l}{L+1}$ — выборочный средний уровень сменяемости p-хабов, $S_p^2 = \frac{(\nu_p - N_p)^2 + \sum_{l=1}^{L} (\nu_p^l - N_p)^2}{L}$ — выборочная несмещенная дисперсия. Критерий выбора устойчивого количества хабов (1.3) будет иметь вид:

$$p_{stab} = arg \min_{p \in \{p_{\min}, \ldots, p_{\max}\}} (N_p - u_{1-\alpha} S_p). \tag{3.8}$$

3.3. Численный эксперимент

В численном эксперименте рассматривается задача нахождения устойчивой сети хабов в формулировке UMApHLP (3.1) — (3.6). Для решения задачи линейного и целочисленного программирования ((3.1)—(3.6)) использовался программный продукт GUROBI Optimizer 7.0.1[1]. Вычисления производились на машине Intel Core i5 2.7GHz с 8GB ОЗУ.

Данные для исследования предоставлены логистической компанией ООО «Деловые Линии», которые содержат данные по 178 грузоприемочным пунктам на территории РФ, множество K содержит 10 объектов, которые потенциально могут быть хабами. Расстояния в часах между объектами сети рассчитаны (без учета пробок) с использованием программного пакета Google Maps Distance Matrix API[2].

В эксперименте $p_{\min} = 6, p_{\max} = 9, R = 40, L = 1000, \alpha = 0.05$ в общей сложности было решено 164 задачи смешанного программирования (англ. Mixed Indeger Programming (MIP)), содержащих 17 890 вещественных переменных, 1 790 бинарных переменных и 21 539 ограничений.

Возмущение спроса производилось путем генерации случайных величин из усеченного нормального распределения.

В Таблице 3.1 представлены частоты сменяемости ν для каждого p до

[1] http://www.gurobi.com/

[2] https://developers.google.com/maps/documentation/distance-matrix/

применения процедуры bootstrap.

p	6	7	8	9
ν	0.875	0.0	0.1	0.625

Таблица 3.1. Частоты сменяемости хабов

В Таблице 3.2 представлены выборочные средние и дисперсии уровней сменяемости для каждого p и значение риска смены хабов (1.4). Наиболее устойчивым количеством хабов является $p = 7$, при котором сеть хабов является постоянной и не изменяется от возмущений в спросе. При $p = 8$ сеть хабов так же является устойчивой, потому что значение VaR близко к 0 и на порядок меньше, чем у $p = 6$ и $p = 9$. Высокий уровень смены хабов в зависимости от возмущений в спросе связан с наличием конкурентных хабов, которые в следствии колебаний в спросе могут быть выгоднее в определенных ситуациях.

p	6	7	8	9
$\bar{N}$	0.87435	0.0	0.10428	0.62343
S^2	0.00265	0.0	0.00245	0.00556
VaR	0.791	0.0	0.014	0.504

Таблица 3.2. Индексы устойчивости VaR

Заключение

В разделе Постановка задачи сформулирована решаемая задача: нахождение устойчивой кластеризации к случайным изменениям данных. В обзоре литературы рассмотрены существующие концепции и методы решения задачи определения «правильной» кластеризации.

В первой главе ВКР формулируются два критерия устойчивости кластеризации. Один из критериев основывается на теории рисков, а именно Value at Risk, где устойчивой кластеризацией считается та, у которой наименьшее значение VaR — индекс отклонений кластеризаций при случайном возмущении исходной выборки данных. Второй критерий основан на оценке ожидания отклонений кластеризаций при многократных имитациях выборок данных случайным образом. Также в Главе 1 представлены процедуры возмущения данных, bootstrap алгоритм генерации множества результатов для принятия статистических решений.

В Главе 2 представлены результаты численного эксперимента на искусственных и реальных выборках данных. Устойчивые кластеризации по обоим критериям совпали во всех расчетах, но концепции критериев отличаются. Полученные результаты сравниваются с тремя индексами устойчивости кластеризации. Описанные в Главе 1 критерии устойчивости получили лучшие результаты на выборках негауссовой структуры, на выборках гауссовой структуры и реальных данных результаты с существующими методами были схожими. Разработана программа для воспроизведения полученных результатов и представлена в открытом доступе.

В Главе 3 критерии устойчивости кластеризации применены к задаче о размещении объектов в сети. Алгоритм определения устойчивой кластеризации адаптирован к нахождению устойчивой сети и количества хабов. Проведен численный эксперимент на реальных данных, где выявлено устойчивое количество хабов в результате решения 164 задач целочисленно-

го программирования. Разработана программа ЭВМ для решения задачи устойчивого размещения хабов и определения устойчивого количества хабов.

Список литературы

1. *Czekanowski J.* Objektive Kriterien in der Ethnologie. — F. Vieweg, 1911.
2. *Jain A. K. et al.* Data clustering: a review // ACM computing surveys (CSUR). — 1999. — Т. 31, № 3. — С. 264—323.
3. *Rokach L., Maimon O.* Clustering methods // Data mining and knowledge discovery handbook. — Springer, 2005. — С. 321—352.
4. *Воронцов К. В.* Лекции по алгоритмам кластеризации и многомерного шкалирования // М.: МГУ. — 2007.
5. *Lozkins A.* Cluster analysis of European countries by an unemployment rate // The XLVI annual international conference on Control Processes and Stability (CPS'15). Abstracts. — St. Petersburg: Publishing House Fedorova G.V., 2015. — С. 111.
6. *Lozkins A., Bure Vladimir M.* The criterion for comparing risks of samples from different distributions // The XLIX annual international conference on Control Processes and Stability (CPS'18). Abstracts. — St. Petersburg: Publishing House Fedorova G.V., 2018. — С. 92.
7. *Ложкинс А., Буре В. М.* Эмпирический подход оценки устойчивости методов кластеризации // Материалы III международной конференции «Устойчивость и процессы управления», посвященная 85-летию со дня рождения профессора, чл.-корр. РАН В. И. Зубова / под ред. А. Жабко, Л. Петросян. — СПб: Издательский Дом Федоровой Г.В., 2015. — С. 431—433.
8. *Ложкинс А., Буре В. М.* Выбор распределительных центров в задаче о размещении объектов на основе процедур статистического моделирования // Материалы XIV международной научной конференции (30

мая — 1 июня 2018г., Москва / под ред. В. Тхай. — М.: ИПУ РАН, 2018. — С. 264—267.

9. *Lozkins A.*, *Bure Vladimir M.* The approach for estimation of clustering robustness // 12th German Probability and Statistics Days. Book of Abstracts. — Deutsche Mathematiker-Vereinigung, 2016. — С. 197—198.

10. *Lozkins A.* Clustering of European Countries by an Inflation Rate and Clusters Research // 20th International Conference of Mathematical Modelling and Analysis. — 2016. — С. 56.

11. *Ложкинс А.*, *Буре В. М.* Критерий сравнения выборок из различных генеральных совокупностей // Процессы управления и устойчивость. — 2018. — С. 475—479.

12. *Ложкинс А.* Кластерный анализ стран Европы по уровню безработицы // Процессы управления и устойчивость. — 2015. — Т. 2, № 1. — С. 641—646.

13. *Lozkins A.*, *Bure V. M.* The method of clusters stability assessing // 2015 International Conference «Stability and Control Processes» in Memory of VI Zubov (SCP). — IEEE. 2015. — С. 479—482.

14. *Старкова Н. В.*, *Ложкинс А.* Кластеризация стран Европы по демографическим признакам // Молодой ученый. — 2016. — № 9. — С. 418—426.

15. *Lozkins A.*, *Bure V. M.* The probabilistic method of finding the local-optimum of clustering // Vestnik Sankt-Peterburgskogo Universiteta. Seriya 10. Prikladnaya Matematika. Informatika. Protsessy Upravleniya. — 2016. — № 1. — С. 28—37.

16. *Lozkins A.* The distribution centres choice in the facility location problem on the basis of statistical modeling procedures // Вестник Санкт-Петер-

бургского университета. Серия 10. Прикладная математика. Информатика. Процессы управления. — 2018. — Т. 14, № 4. — С. 346—351.

17. *Lozkins A.*, *Bure V. M.* Single hub location-allocation problem under robustness clustering concept // Вестник Санкт-Петербургского университета. Серия 10. Прикладная математика. Информатика. Процессы управления. — 2017. — Т. 13, № 4. — С. 398—406.

18. *Lozkins A.* Stability-based approach of cluster number determination : дис. ... маг. / Lozkins Aleksejs. — St. Petersburg State University, 2016.

19. *Ложкинс А.*, *Буре В. М.* Программа для определения устойчивого количества распределительных центров. — 2018. — Свидетельство о государственной регистрации программы для ЭВМ No.2018665042 от 29.11.2018.

20. *Ben-Hur A.*, *Elisseeff A.*, *Guyon I.* A stability based method for discovering structure in clustered data // Biocomputing 2002. — World Scientific, 2001. — С. 6—17.

21. *Fowlkes E. B.*, *Mallows C. L.* A method for comparing two hierarchical clusterings // Journal of the American statistical association. — 1983. — Т. 78, № 383. — С. 553—569.

22. *Ochiai A.* Zoogeographical studies on the soleoid fishes found in Japan and its neighbouring regions-I // Bull. Jpn. Soc. scient. Fish. — 1957. — Т. 22. — С. 522—525.

23. *Jaccard P.* Distribution de la flore alpine dans le bassin des Dranses et dans quelques régions voisines // Bull Soc Vaudoise Sci Nat. — 1901. — Т. 37. — С. 241—272.

24. *Jain A. et al.* Algorithms for clustering data. Т. 6. — Prentice hall Englewood Cliffs, 1988.

25. *Peterson J. D.* Clustering overview. — 2002.

26. *Нейский И. М.* Классификация и сравнение методов кластеризации // Интеллектуальные технологии и системы. Сб. учебно-методических работ и статей аспирантов и студентов. М.: НОК «CLAIM. — 2006. — № 8. — С. 130—142.

27. *Marriott F.* Practical problems in a method of cluster analysis // Biometrics. — 1971. — С. 501—514.

28. *Caliński T.*, *Harabasz J.* A dendrite method for cluster analysis // Communications in Statistics-theory and Methods. — 1974. — Т. 3, № 1. — С. 1—27.

29. *Krzanowski W. J.*, *Lai Y.* A criterion for determining the number of groups in a data set using sum-of-squares clustering // Biometrics. — 1988. — С. 23—34.

30. *Hartigan J. A.* Clustering algorithms. — 1975.

31. *Davies D. L.*, *Bouldin D. W.* A cluster separation measure // IEEE transactions on pattern analysis and machine intelligence. — 1979. — № 2. — С. 224—227.

32. *Dunn J. C.* Well-separated clusters and optimal fuzzy partitions // Journal of cybernetics. — 1974. — Т. 4, № 1. — С. 95—104.

33. *Граничин О. Н. и др.* Рандомизированный алгоритм нахождения количества кластеров // Автоматика и телемеханика. — 2011. — № 4. — С. 86—98.

34. *Шалымов Д. С.* Рандомизированный метод определения количества кластеров на множестве данных // Научно-технический вестник информационных технологий, механики и оптики. — 2009. — 5 (63).

35. *Goutte C. et al.* Feature-space clustering for fMRI meta-analysis // Human brain mapping. — 2001. — Т. 13, № 3. — С. 165—183.

36. *Akaike H.* A new look at the statistical model identification // Selected Papers of Hirotugu Akaike. — Springer, 1974. — С. 215—222.

37. *Schwarz G.* Estimating the dimension of a model Ann Stat 6: 461–464 // Find this article online. — 1978.

38. *Biernacki C.*, *Celeux G.*, *Govaert G.* Assessing a mixture model for clustering with the integrated completed likelihood // IEEE transactions on pattern analysis and machine intelligence. — 2000. — Т. 22, № 7. — С. 719—725.

39. *Sugar C. A.*, *James G. M.* Finding the number of clusters in a dataset: An information-theoretic approach // Journal of the American Statistical Association. — 2003. — Т. 98, № 463. — С. 750—763.

40. *Ng A.* Clustering with the k-means algorithm // Machine Learning. — 2012.

41. *Hennig C.* Cluster-wise assessment of cluster stability // Computational Statistics & Data Analysis. — 2007. — Т. 52, № 1. — С. 258—271.

42. *Shamir O.*, *Tishby N.* Cluster stability for finite samples // Advances in neural information processing systems. — 2008. — С. 1297—1304.

43. *Ben-Hur A.*, *Guyon I.* Detecting stable clusters using principal component analysis // Functional genomics. — Springer, 2003. — С. 159—182.

44. *Tibshirani R.*, *Walther G.* Cluster validation by prediction strength // Journal of Computational and Graphical Statistics. — 2005. — Т. 14, № 3. — С. 511—528.

45. *Pascual D.*, *Pla F.*, *Sánchez J. S.* Cluster validation using information stability measures // Pattern Recognition Letters. — 2010. — Т. 31, № 6. — С. 454—461.

46. *Firdaus S.*, *Uddin M. A.* A survey on clustering algorithms and complexity analysis // International Journal of Computer Science Issues (IJCSI). — 2015. — Т. 12, № 2. — С. 62.

47. *Efron B.*, *Tibshirani R. J.* An introduction to the bootstrap. — CRC press, 1994.

48. *Artzner P. et al.* Coherent measures of risk // Mathematical finance. — 1999. — Т. 9, № 3. — С. 203—228.

49. *Rockafellar R. T. et al.* Optimization of conditional value-at-risk // Journal of risk. — 2000. — Т. 2. — С. 21—42.

50. *Ahmadi-Javid A.* Entropic value-at-risk: A new coherent risk measure // Journal of Optimization Theory and Applications. — 2012. — Т. 155, № 3. — С. 1105—1123.

51. *Буре В.*, *Парилина Е.* Теория вероятностей и математическая статистика // СПб.: Лань. — 2013. — Т. 416.

52. *Littlewood K.* Forecasting and control of passenger bookings // Airline Group International Federation of Operational Research Societies Proceedings, 1972. — 1972. — Т. 12. — С. 95—117.

53. *Asuncion A.*, *Newman D.* UCI machine learning repository. — 2007.

54. *O'kelly M. E.* The location of interacting hub facilities // Transportation science. — 1986. — Т. 20, № 2. — С. 92—106.

55. *Contreras I.* Hub location problems // Location science. — Springer, 2015. — С. 311—344.

56. *Alumur S. A.*, *Nickel S.*, *Saldanha-da-Gama F.* Hub location under uncertainty // Transportation Research Part B: Methodological. — 2012. — Т. 46, № 4. — С. 529—543.

57. *Merakli M.*, *Yaman H.* Robust intermodal hub location under polyhedral demand uncertainty // Transportation Research Part B: Methodological. — 2016. — T. 86. — C. 66—85.

58. *Contreras I.*, *Cordeau J.-F.*, *Laporte G.* Stochastic uncapacitated hub location // European Journal of Operational Research. — 2011. — T. 212, № 3. — C. 518—528.

59. *Hamacher H. W. et al.* Adapting polyhedral properties from facility to hub location problems // Discrete Applied Mathematics. — 2004. — T. 145, № 1. — C. 104—116.