Методика адаптивной кластеризации фактографических данных на основе интеграции методов минимального остовного дерева и нечетких К-средних

Нейский И.М.

Научный руководитель:
к.т.н., доцент, Филиппович А.Ю.

Задачи интеллектуального анализа данных

Кластеризация – разбиение совокупности объектов на однородные группы (кластеры)

Проблемы кластерного анализа

1. Выбор метода исследования [И.Д. Мандель]	3. Выбор значения параметра «Количество кластеров» [А.А. Баргесян, М.С. Куприянов, В.В. Степаненко, И.И. Холод, Н.Б. Паклин]
2. Оценка качества полученного разбиения [Н.Б. Паклин, A. K. Jain, R.C. Dubes, Б. Дюран, П. Оделл]	4. Постоянно увеличивающиеся объемы данных

Адаптивная кластеризация – кластеризация, для которой количество кластеров является результатом исследования

	MST	**Fuzzy C-means**
Назначение	Кластеризация больших наборов числовых данных	Кластеризация больших наборов числовых данных
Достоинства	Выделяет кластеры произвольной формы, в т.ч. кластеры выпуклой и вогнутой	Плавное разбиение объектов
Недостатки	Работа только с числовыми данными, четкое распределение объектов	Вычислительная сложность, задание количества кластеров, возникает неопределенность с объектами, которые удалены от центров всех кластеров, выделение кластеров сферической формы

Примеры практических задач

Показатель	Количественная характеристика	Качественная характеристика
Количество исходных объектов (шт.)	500 - 50 000	-
Количество значимых характеристик объектов (шт.)	70 - 150	-
Типы характеристик	два вида	числовые, лингвистические
Форма получаемых кластеров	-	сложная, с пересечениями
Количество получаемых кластеров (шт.)	5 - 30	результат исследования

— Разработка тарифной сетки;

— Группировка перечня предлагаемых услуг;

— Формирование потребительской корзины;

— Сегментирование сферы деятельности;

— Распознавание изображений;

— Тематический анализ библиотеки документов;

— Оптимизация использования складских помещений;

— Выделение потенциальных покупателей;

— Построение показательной выборки и др.

Цель и задачи

Цель

Разработка методики адаптивной кластеризации фактографических данных на основе интеграции методов минимального остовного дерева и нечетких К-средних

Назначение

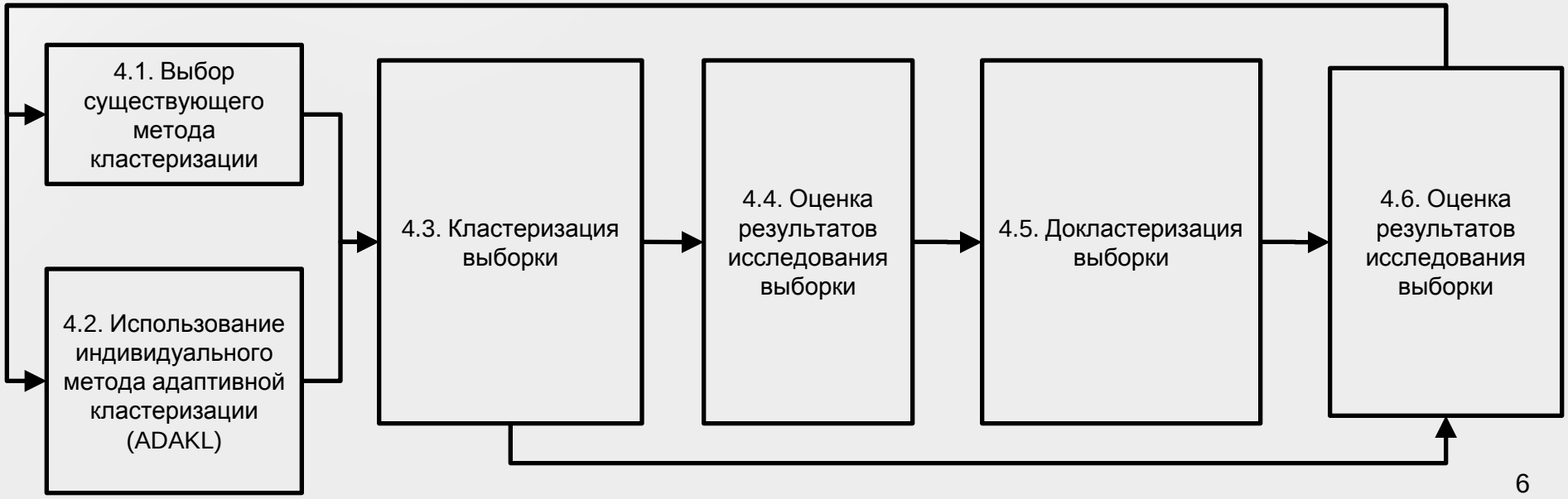
Предназначена для аналитиков, специалистов по анализу данных, разработчиков систем класса Data Mining

Задачи

1. Исследование методов и подходов интеллектуального анализа данных, используемых для кластеризации фактографических данных.
2. Разработка методики адаптивной кластеризации фактографических данных.
3. Разработка рекомендаций по выбору существующих методов кластеризации.
4. Разработка метода адаптивной кластеризации фактографических данных.
5. Разработка метода докластеризации.
6. Разработка программного комплекса для автоматизации этапов метода кластеризации.
7. Оценка эффективности разработанного метода с помощью экспериментальных исследований.

Методика адаптивной кластеризации

Выбор метода кластерного анализа



Выбор существующего метода кластеризации

Шаг 1. Выбор метода кластерного анализа

1. На основе существующих рекомендаций по исследованию предметных областей и задач.
2. На основе критериев
3. Общий алгоритм

Шаг 2. Настройка параметров выбранного метода кластерного анализа

— объем обучающего множества;
— объем валидационного множества;
— объем тестового множества;
— количество атрибутов входного набора данных;
— тип атрибутов входного набора данных;
— используемость атрибутов входного набора данных.

Характеристические параметры

— количество кластеров;
— алгоритм выполнения дополнительной кластеризации;
— пороговое значение остановки работы алгоритма;
— способ выбора начальных центров;
— максимальное количество итераций;
— количество одновременно обрабатываемых данных;
— количество предварительных разделов;
— коэффициент удаленности.

Итерационные параметры

— способ определения расстояния между кластерами;
— метод оценки качества кластеризации;
— пороговое значение для метода оценки качества кластеризации;
— начальное пороговое значение алгоритма;
— процент аномалий (выбросов) в полном объеме;
— разделяющая функция;
— скорость обучения сети.

Экспертные параметры

Шаг 3. Анализ массива фактографических данных и оценка разбиения

Кластеризация массива

Аналитическая оценка

Индекс оценки$_1$

Индекс оценки$_2$

…

Индекс оценки$_к$

Качество?

Да

Результат

Нет

Выбор метода кластеризации на основе рекомендаций

Метод	CURE	BIRCH	MST	k-средних	PAM	CLOPE	SOM	HCM	Fuzzy C-Means
Рекомен-дации	Выявляет кластеры произвольной формы, метод менее чувствителен к выбросам, чем MST. Время работы алгоритма незначительное	Метод предназначен для очень больших наборов данных. Работает с произвольным количеством оперативной памяти. Получаемое разбиение обладает высоким качеством	Лучше всего подходит для выделения кластеров произвольной формы	Показывает хорошие результаты при работе с данными, которые распределены по компактным группам сферической формы	Показывает хорошие результаты при работе с данными, которые распределены по компактным группам сферической формы	Кластеризация больших объемов категорийных данных	Поиск и анализ закономер-ностей	Показывает высокие результаты при работе с данными, которые распределены по компактным группам сферической формы	Относит объект к разным кластерам на основе степени принадлежности элемента к кластерам, выделяет кластеры сферической формы
Противо-показания	Не использовать для исследования объектов с большим количеством атрибутов, требует задания пороговых значений	Не использовать для получения несферических форм кластеров, требует задания пороговых значений	Очень чувствителен к выбросам и может медленно работать на больших массивах данных	Очень чувствителен к выбросам и может медленно работать на больших массивах данных	Чувствителен к выбросам и может медленно работать на больших массивах данных	Требуется подбор оптимального коэффициента отталкивания	Требуется минимизация размеров карты, проблема с аналитическим обоснованием результатов исследования	Чувствителен к выбросам и может медленно работать на больших массивах данных	Высокие требования к вычислительной мощности используемого аппаратного обеспечения, не работает с объектами, которые удалены от всех кластеров, и с вложенными кластерами

Критерии выбора методов кластеризации

Критерий 1 (Cr_1)	*Объем информации*
Критерий 2 (Cr_2)	*Размерность информации*
Критерий 3 (Cr_3)	*Типы атрибутов*
Критерий 4 (Cr_4)	*Чувствительность к равномерности информации*
Критерий 5 (Cr_5)	*Априорное (экспертное) представление о форме получаемых кластеров*
Критерий 6 (Cr_6)	*Априорное (экспертное) представление о количестве кластеров*
Критерий 7 (Cr_7)	*Априорное (экспертное) представление о перекрываемости кластеров*
Критерий 8 (Cr_8)	*Качество кластеризации*

Выбор метода кластеризации на основе критериев

Если Критерий 6 = «Неизвестно» и Критерий 8 <> «Высокое»

То Выдать сообщение «Метод кластеризации отсутствует»

Или Если Критерий 8 = «Низкое» и Критерий 7 = «Без пересечений» и Критерий 5 = «Сферическая форма» и Критерий 4 = «Данные равномерны» и Критерий 3 = «Числовые атрибуты» и Критерий 2 = «Низкая размерность»

То Метод := «k-средних»

Или если Критерий 8 = «Среднее»

То

Начало блока 1

Если Критерий 7 = «С пересечениями» и Критерий 5 = «Сложная форма» и Критерий 4 = «Данные равномерны» и Критерий 3 = «Числовые атрибуты» и Критерий 2 <= «Высокая размерность»

То Метод := Fuzzy C-means

Или если Критерий 7 = «Без пересечений» и Критерий 5 = «Сферическая форма» и Критерий 4 = «Данные равномерны» и Критерий 3 = «Числовые атрибуты» и Критерий 2 = «Невысокая размерность»

То Метод := BIRCH

Или если Критерий 7 = «Без пересечений» и Критерий 5 = «Сложная форма» и Критерий 4 = «Данные неравномерны» и Критерий 3 = «Числовые атрибуты» и Критерий 2 <= «Высокая размерность»

То Метод := Самоорганизующиеся карты Кохонена

Или если Критерий 7 = «Без пересечений» и Критерий 5 = «Сложная форма» и Критерий 4 = «Данные равномерны» и Критерий 3 = «Числовые атрибуты» и Критерий 2 <= «Средняя размерность» То

То Метод := HCM

Или если Критерий 7 = «Без пересечений» и Критерий 5 = «Сложная форма» и Критерий 4 = «Данные равномерны» и Критерий 3 = «Числовые атрибуты» и Критерий 2 = «Низкая размерность» и Критерий 1 = «Небольшой объем данных»

То Метод := PAM

Или

Выдать сообщение «Метод кластеризации отсутствует»

Конец блока 1

Или если Критерий 8 = «Высокое»

То

Начало блока 2

Если Критерий 7 = «Без пересечений» и Критерий 5 = «Сложная форма» и Критерий 4 = «Данные неравномерны» и Критерий 3 = «Числовые атрибуты» и Критерий 2 = «Низкая размерность»

То Метод := CURE

Или Если Критерий 7 = «Без пересечений» и Критерий 5 = «Сложная форма» и Критерий 4 = «Данные равномерны» и Критерий 3 = «Числовые атрибуты» и Критерий 2 <= «Средняя размерность» и Критерий 1 <= «Большой объем данных»

То Метод := MST

Или Если Критерий 7 = «Без пересечений» и Критерий 5 = «Сложная форма» и Критерий 4 = «Данные равномерны» и Критерий 3 = «Числовые атрибуты» и Критерий 2 <= «Средняя размерность» и Критерий 1 <= «Огромный объем данных» То

То Метод := CLOPE

Или

Выдать сообщение «Метод кластеризации отсутствует»

Конец блока 2

Или

Выдать сообщение «Метод кластеризации отсутствует»

Метод адаптивной кластеризации ADAKL

[Dunn J. C., Bezdek J. C., Загоруйко Н.Г., Елкина В.Н.]

Описание метода ADAKL

Оценка информационного расстояния между объектами

$$Dist_{ij} = \left\| u_i - u_j \right\| = Metric\left(u_i, u_j\right)$$

$Metric = Åâêëèäîâ\ âî\ \ ðàññòîÿíèå$

$$Dist_{ij} = \sqrt{\sum_l \left(\left[Value_{il} - Value_{jl} \right] * K_l \right)^2}$$

$Metric = Êâàäðàò\ Åâêëèäîâà\ ðàññòîÿíèÿ$

$$Dist_{ij} = \sum_l \left(\left[Value_{il} - Value_{jl} \right] * K_l \right)^2$$

$Metric = Ðàññòîÿíèå×ááû ø åâà$

$$Dist_{ij} = \underset{l}{Max} \left[\left(Value_{il} - Value_{jl} \right) * K_l \right]$$

$$K = \left\{ K_1, K_2, ..., K_n \right\} \quad K_i \in [0;1]$$

Первичное разделение на кластеры

$$Dist_{ij}^{l} := \left\{ 0 \left| Dist_{ij}^{l} = Max, i,\ j = \overline{1,m}, l = \overline{1,k} \right. \right\}$$

Расчет центров кластеров

$$V_i^k = Avg\left(\left\{ u_j \left| u_j \in C_i^k \right. \right\} \right)$$

$$Avg[r] = \frac{\sum_{u_j \in V_i^k} \left\{ Value_{jr} \left| FieldType[w] = "Âõî äí î å" \right. \right\}}{\left| C_i^k \right|} \quad \begin{matrix} i = \overline{1,k} & j = \overline{1,m} \\ r = \overline{1,n} & l = \overline{1,m} \end{matrix}$$

$$Avg[r] = \left\{ \underset{\sum \left\| Value_{jr} - Value_{lr} \right\| = min\ èëè \quad \varphi = max}{Value_{jr}} \left| FieldType[w] = "Âõî äí î å" \right. \right\}$$
$$u_j, u_l \in C_i^k \qquad u_j, u_l \in C_i^k$$

Отнесение объектов к кластерам

$$u_j \in V_i^k \left| Dist_{ij}^{k'} \le w\ èëè\ Dist_{ij}^{k'} = \underset{i}{Min}\left(Dist_{ij}^{k'} \right) \right. \quad i = \overline{1,k} \quad j = \overline{1,m}$$

Расчет центров кластеров (2)

$$V_i^{k'} = \frac{\sum_{j=1}^{m} \mu_{ij}^{\ p} * u_j}{\sum_{j=1}^{m} \mu_{ij}^{\ p}}, i = \overline{1,k}$$
$$u_j \in V_i^{k'}$$

$$V_i^{k'}[r] = \underset{\mu_{ij} = Max\left(\mu_{lj}\right)}{Value_{jr}}$$

Степень принадлежности

$$\mu_{ij} = \left(1 - Dist_{ij}^{k'}\right)^2 \quad i = \overline{1,k} \quad j = \overline{1,m}$$

Выбор лучшего разбиения

$$O^k = \frac{\sum_{i=\overline{1,k}} \left|V_i^{k'}\right| * \sum_{j=1}^{m} \mu^p{}_{ij} * \left\|V_i^{k'} - u_j\right\|}{\underset{i \neq j}{Min}\left(\left\|V_i^{k'} - u_j\right\|\right) * Ma\underset{u_j \in V_i^{k'}}{x}\left(\left\|V_i^{k'} - u_j\right\|\right) * \sum_{j=1}^{m}\left\|V_i^{k'} - u_j\right\| * k}$$

1

3

2 $m * k^2$

$$O_{\hat{I}\ i\ \grave{o}} = \underset{i=\overline{1,q}}{MAX}\left(O^i\right)$$

Аналитическая оценка сложности метода

Алгоритм построения MST / Способ нормализации	Алгоритм Борувки	Алгоритм Крускала	Алгоритм Прима
Линейная нормализация	1.1 $O\left(m^2 * \left[a+b+Lg\left(m\right)+q\right]\right)$	2.1 $O\left(m^2 * \left[a+b+Lg\left(m^2\right)+q\right]\right)$	3.1 $O\left(m^2 * \left[a+b+Lg\left(m\right)+q\right]\right)$
Статистическая нормализация	1.2 $O\left(m^2 * \left[a+b+Lg\left(m\right)+q\right]\right)$	2.2 $O\left(m^2 * \left[a+b+Lg\left(m^2\right)+q\right]\right)$	3.2 $O\left(m^2 * \left[a+b+Lg\left(m\right)+q\right]\right)$

Итоговая аналитическая оценка метода

$$O\left(m^2 * \left[a+b+Lg\left(m\right)+q\right]\right) \; ; \; O\left(m^2 * \left[a+b+Lg\left(m^2\right)+q\right]\right)$$

1.1

2.1

3.1

1.2

2.2

3.2

Исследуемые наборы данных

Характеристика / Исследование	Общее количество исходных данных (шт. записей)	Количество атрибутов числового типа [количество значимых атрибутов] (шт.)	Количество атрибутов текстового атрибута [количество значимых атрибутов] (шт.)	Общее количество атрибутов [общее количество значимых атрибутов] (шт.)
1	267	0 [0]	5 [2]	5 [2]
2	533	73 [72]	3 [3]	76 [75]
3	267	5 [2]	4 [2]	9 [4]
4	450	3 [3]	2 [1]	5 [4]

SecBrief	SecType	SecBranch	Nominal	Volume
46014	Облигации	Госзайм	1 000.00	10 000 000 000.00
46018	Облигации	Госзайм	1 000.00	10 000 000.00
48001	Облигации	Госзайм	1 000.00	24 099 483.00
DCL ASSETS Int. Ltd.	Облигации	ЗапФинСектор	100 000.00	1 380.00
IMPEX07	Облигации	ЗапФинСектор	2 000.00	100 000 000.00
RusFed05	Облигации	ВнешГосзайм	1 000.00	1 000 000.00
RusFed18	Облигации	ВнешГосзайм	1 000.00	1 000 000.00
RusFed30	Облигации	ВнешГосзайм	1.00	1 000 000.00
SISFIN08	Облигации	ЗапФинСектор	1 000.00	350 000.00
ГМК Норникель АО в.5	Акции	Металлургия	1.00	70 000 000.00
Газэнергопромбанк АОИ	Акции	Банк	1.00	100 000 000.00
Казаньоргсинтез АПИ 2	Акции	ХимПром	1.00	25 000 000.00
Конкордия-АВАНТАЖ	Акции	ЖилСектор	0.00	3 000.00

Vehicle	TopSpeed	Colour	AirResistance	Weight	Type
V1	220	red	0.30	1 300.00	Sport
V2	230	black	0.32	1 400.00	Sport
V3	260	red	0.29	1 500.00	Sport
V4	140	grey	0.35	800.00	Medium market
V5	155	blue	0.33	950.00	Medium market
V6	130	white	0.40	600.00	Medium market
V7	100	black	0.5	3000	Lorry
V8	105	red	0.60	2 500.00	Lorry
V9	110	grey	0.55	3 500.00	Lorry

ClientID	ClientType	BuyCnt	SellCnt	BuyTurn	SellTurn	Freq	FinInstrCnt	SecGroupType
Клиент1 - 1	Долгосрочный инвестор	2	0	2000	0	0.001	3	Акции
Клиент2 - 2	Спекулянт	10	10	150	200	1	9	Акции
Клиент3 - 3	Потребитель	0	3	0	30	0.01	2	Акции
Клиент4 - 4	Паевой фонд	25	25	28000	32000	1	15	Акции
Клиент5 - 5	Акционер	1	0	8000	0	0.00001	1	Акции
Клиент6 - 1	Долгосрочный инвестор	3	0	1355	0	0.003	2	Акции
Клиент7 - 2	Спекулянт	18	8	570	560	0.999	5	Акции,

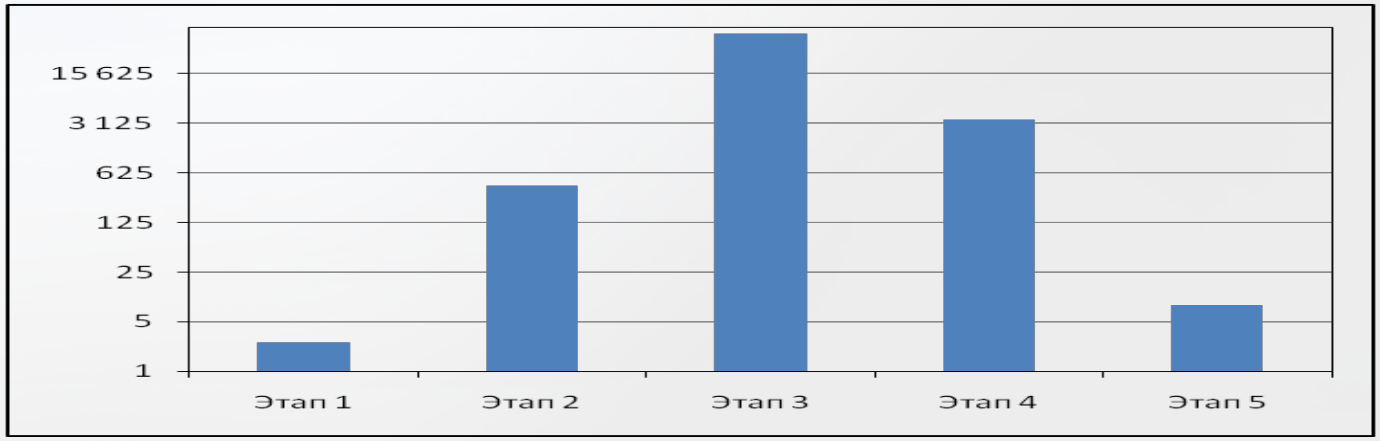
Экспериментальная оценка метода адаптивной кластеризации

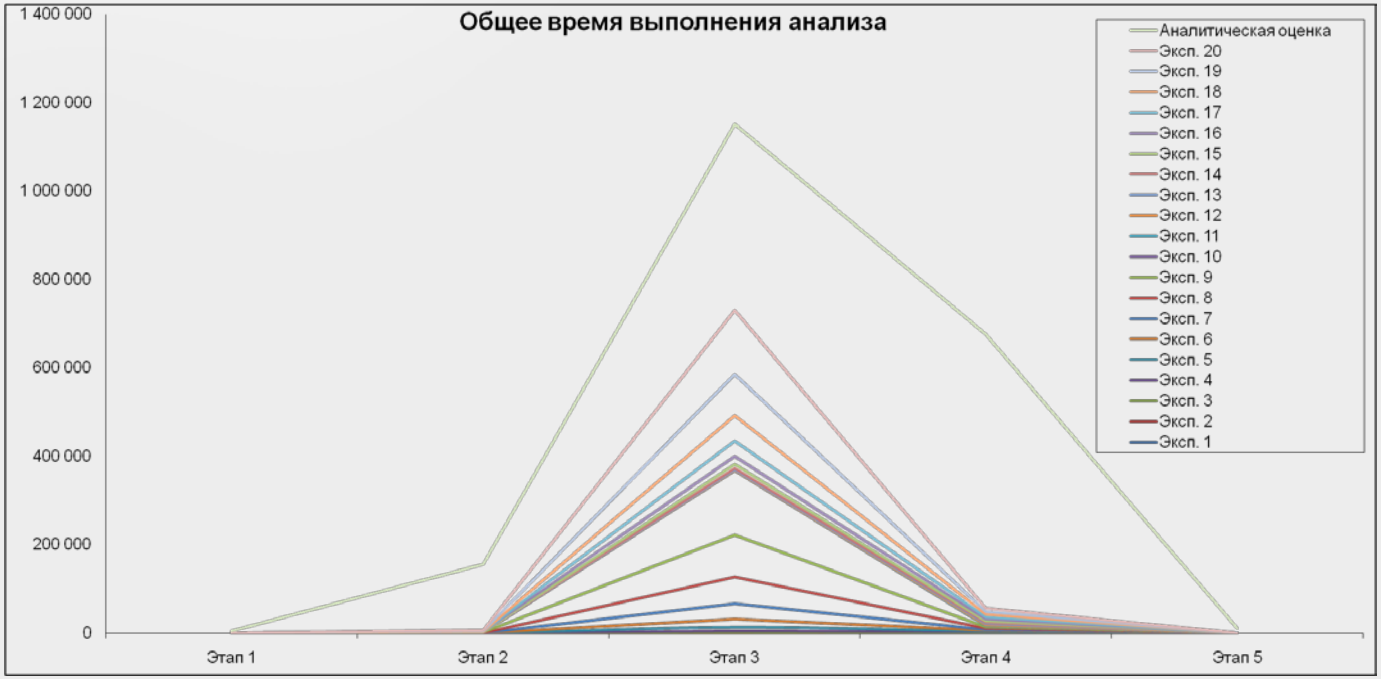
Исследование 1: выделение секторов инвестирования на основе анализа показателей финансовых инструментов;
Исследование 2: выделение групп клиентов на основе статистических данных о деятельности клиентов за период;
Исследование 3: выявление категорий финансовых инструментов для оценки эффективности операций;
Исследование 4: выделение классов автомобилей на основе данных об их характеристиках.

Показатель / Оценка	Средневзвешенная оценка разбиения	Средневзвешенная оценка разбиения с заданным количеством кластеров (без учета лингвистических атрибутов)	Средневзвешенная оценка разбиения с заданным количеством кластеров (с учетом лингвистических атрибутов)	Итоговая оценка
Метод 1	0.7913	0.9150	0.9237	0.8767
Метод 2	-	0.8232	-	0.8232
Метод 3	0.9762	0.9981	0.9990	0.9911

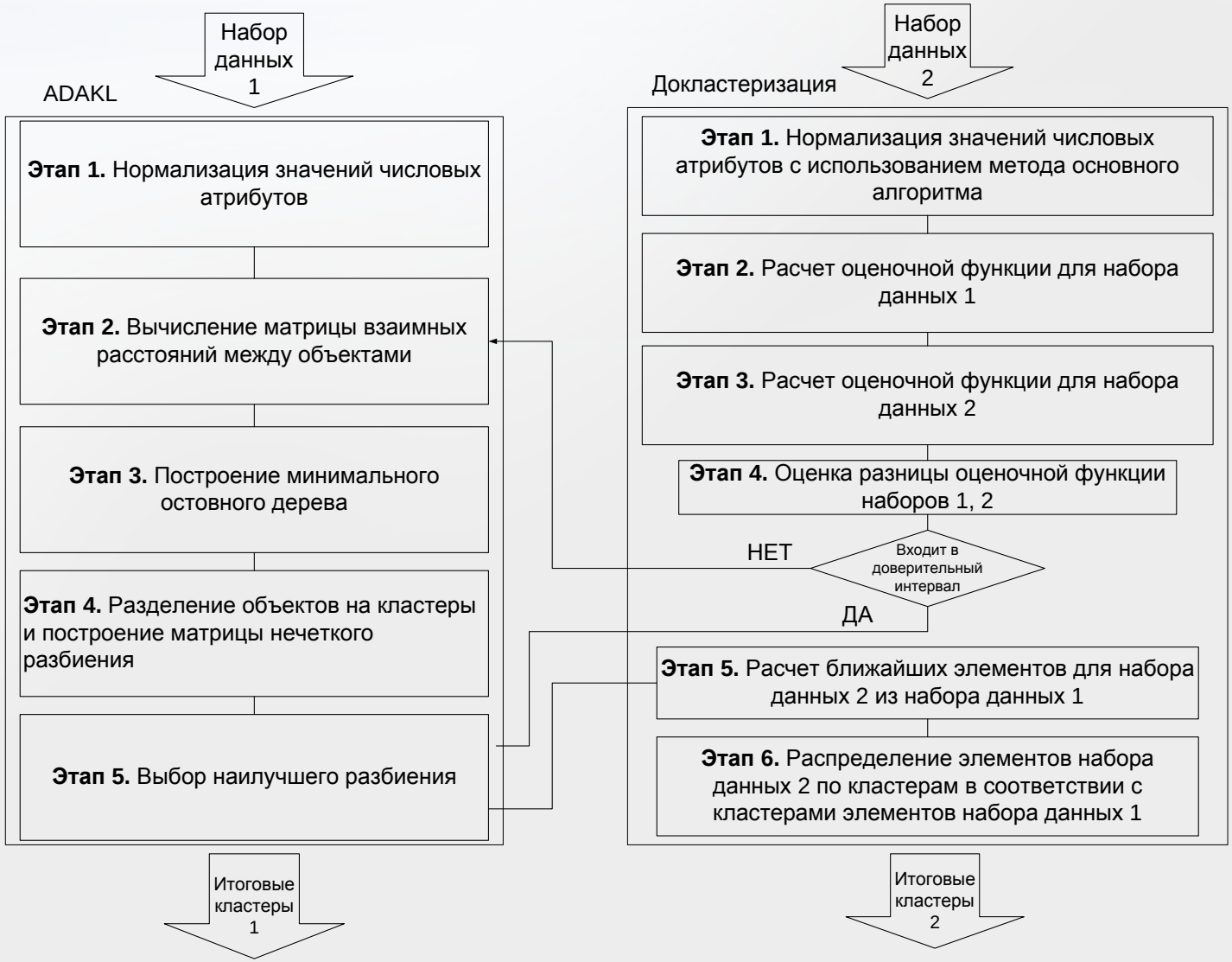
Метод 1 – самоорганизующиеся карты Кохонена
Метод 2 – метод k-средних
Метод 3 – метод ADAKL

Общее время анализа по этапам





Метод докластеризации



Оценка исходных наборов данных

$$D_1 = \sqrt{\frac{\sum_{i=1}^{r} \left\| A_i - Avg[A] \right\|^2}{r}}$$

$$D_2 = \sqrt{\frac{\sum_{i=1}^{o} \left\| B_i - Avg[B] \right\|^2}{o}}$$

$$\left| D_1 - D_2 \right| \le \delta$$

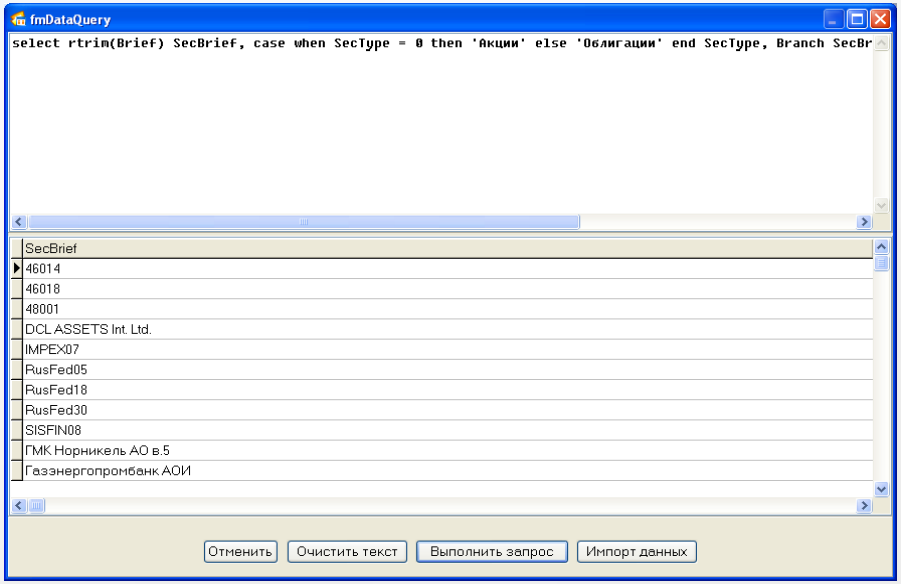
Распределение объектов по кластерам

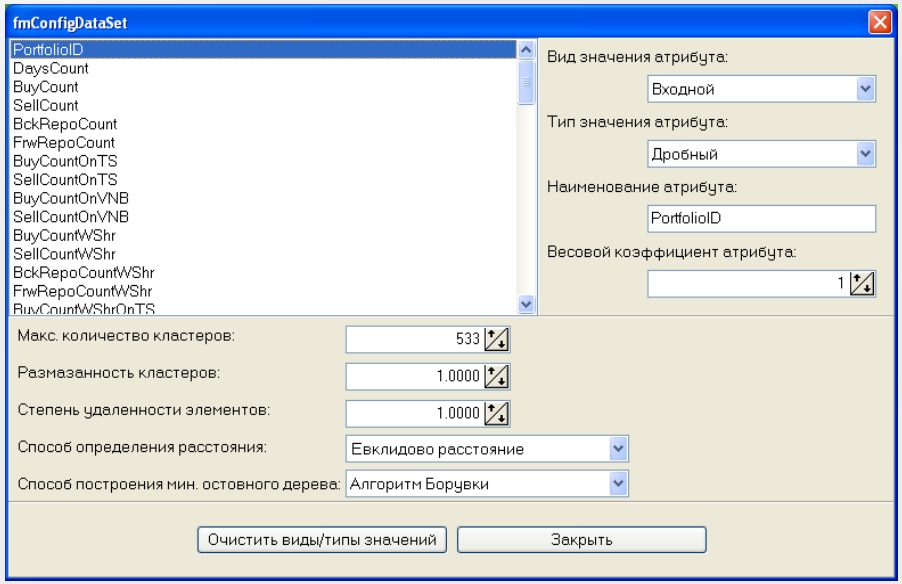
$$\mu_{ij} = \left(1 - Dist_j^{Norm}\right) * \mu_{il} =$$

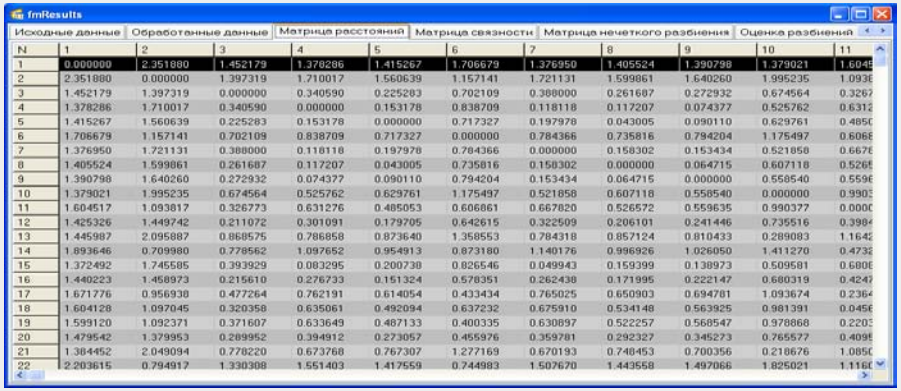
$$\left[1 - Dist_{lj} \Big/ Max\left(Dist_{sj}\right)\right] * \mu_{il}$$

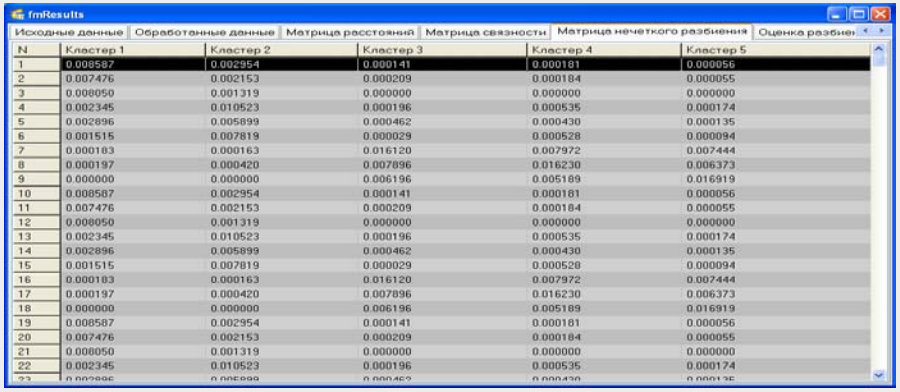
Программное решение (1)

Программное решение (2)

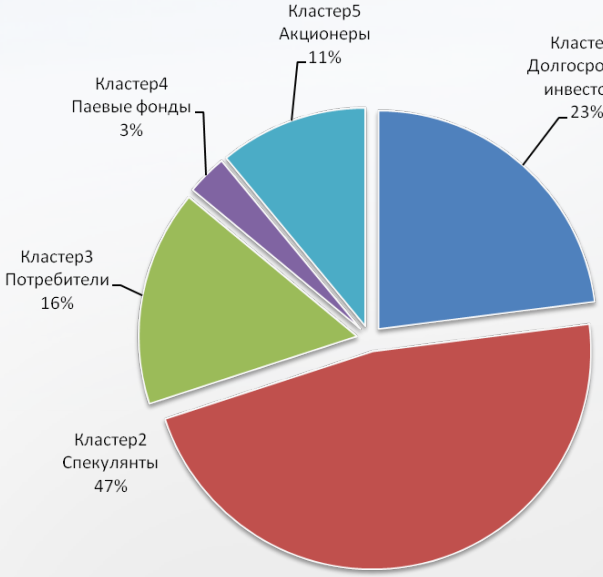


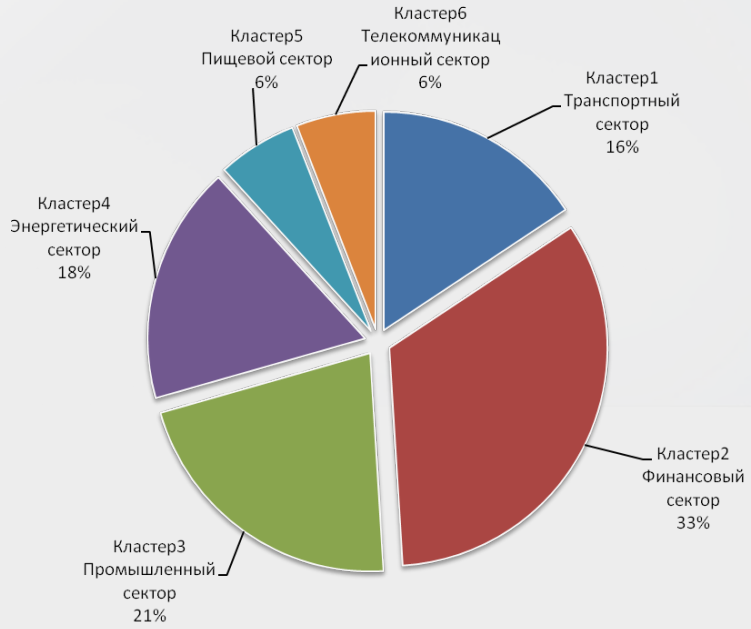


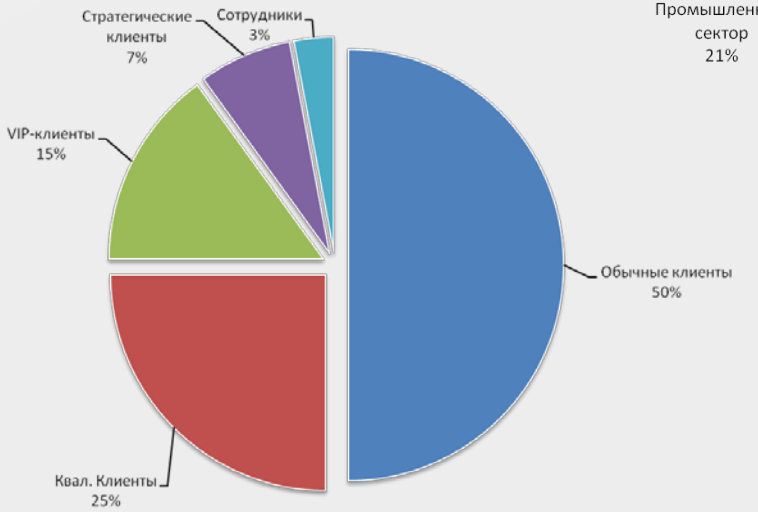




Практическое исследование







Основные результаты работы

1. Проведено исследование существующих методов и подходов интеллектуального анализа данных, используемых для кластеризации фактографических данных.
2. Проведен анализ аналитических программных комплексов с выделением назначения программного комплекса и основных функциональных возможностей.
3. Разработана общая методика адаптивной кластеризации, которая состоит из пяти этапов: выборка исходных данных, исследование полученной выборки с целью выявления значимых для разбиения характеристик, разработка контрольного примера, выбор метода кластеризации, кластеризации полного объема данных.
4. Для выбора метода кластеризации на основе литературных источников выделено восемь критериев.
5. Разработан критерий для оценки качества разбиения, который позволяет проводить оценку и сравнение результатов исследований на основе сравнения итоговых и ожидаемых количественных показателей разбиения.
6. Разработан метод адаптивной кластеризации (ADAKL) на основе интеграции методов минимального остовного дерева и нечетких К-средних, определяющий количество кластеров на основе локального критерия, обладающий двухэтапностью, восемью входными параметрами настройки, нечеткостью при распределении объектов по кластерам, возможностью использования объектов с разными типами атрибутов, приемлемым временем работы и конечностью результата.
7. Разработан локальный критерий оценки разбиения множества на кластеры, который учитывает характеристики практической задачи, лежащей в основе научного исследования: выделение кластеров с наименьшими взаимными расстояниями и наибольшим количеством элементов в кластере, минимизация количества кластеров, минимизация взаимных расстояний между получаемыми центрами кластеров и распределяемыми объектами.
8. Разработан метод докластеризации, позволяющий расширять исследованные массивы фактографических данных и уменьшающий затраты времени на проведение исследования за счет выявления взаимных связей между исследованными объектами и добавляемыми объектами.
9. Разработанный метод ADAKL реализован в виде программного решения, который подтверждает аналитическую оценку.
10. На основе программного решения проведены экспериментальные исследования и оценка состоятельности разработанного метода в сравнении с имеющимися методами (k – средние, карты Кохонена).