Нейский Иван Михайлович

Методика адаптивной кластеризации фактографических данных на базе Fuzzy C-means и MST

05.13.17 – Теоретические основы информатики

АВТОРЕФЕРАТ
диссертации на соискание ученой степени
кандидата технических наук

Научный руководитель:
кандидат технических наук,
доцент Филиппович А.Ю.

Москва
2010

Работа выполнена на кафедре «Системы обработки информации и управления» Московского государственного технического университета им. Н.Э. Баумана

Научный руководитель: кандидат технических наук, доцент Филиппович Андрей Юрьевич

Официальные оппоненты:

Ведущая организация:

Защита диссертации состоится «__» _____ 2010 г. в __:__ на заседании диссертационного совета Д 212.147.03 при Московском государственном университете печати по адресу: 127550, Москва, ул. Прянишникова, 2а.

С диссертацией можно ознакомиться в библиотеке Московского государственного университета печати. Автореферат разослан «__» _____ 2010 г.

Ученый секретарь диссертационного совета Иващенко И.М.

Общая характеристика работы

Актуальность работы

Использование корпоративных информационных систем ведет к росту объемов информации. Широкое применение этих систем повышает необходимость в использовании аналитических систем вместо человеческих ресурсов для извлечения знаний из накопленной информации, делая актуальной задачу разработки специализированных методик и программных инструментов.

Изучением проблем и созданием решений в этой области активно занимаются направления Business Intelligence (Интеллектуальный анализ данных) и Knowledge Management (Управление знаниями), в рамках которых выделяются поднаправления Knowledge Discovery in Databases (Выявление знаний в базах данных), Data Mining (Анализ фактографических данных), Text Mining (Анализ неструктурированных данных) и др. Результаты исследований этих направлений положены в основу многих информационно-аналитических систем, которые используются, в основном, для персональной работы экспертов. Однако, современной тенденцией является применение указанных технологий и для централизованного управления организациями.

Для исследования структурированных массивов информации используется анализ фактографических данных, в котором выделены шесть различных задач, такие как: классификация, регрессия, кластеризация, выявление ассоциаций, выявление последовательностей, прогнозирование.

Потребность в кластеризации возникает в тех областях/этапах деятельности, где есть необходимость в разбиении объектов (ситуаций) на непересекающиеся подмножества, называемыми кластерами, так, чтобы каждый кластер состоял из схожих объектов, а объекты разных кластеров существенно отличались. Четкое разделение на кластеры возможно только в идеальных условиях и при сильно различающихся параметрах объектов кластеризации. Поэтому для решения реальных задач все чаще применяются нечеткие методы, в которых разбиение объектов (ситуаций) выполняется на частично пересекающиеся подмножества.

Важной предпосылкой применения нечетких методик кластеризации в реальных условиях является то, что характеристики объектов не всегда являются измеримыми и поэтому в ряде случаев присутствуют экспертные оценки характеристик объектов, которые являются субъективными и могут быть противоречивыми.

В связи с ростом динамики изменений в социально-экономической и научно-производственной среде задача кластеризации актуальна в различных сферах и предметных областях, например: выделение групп клиентов брокерского обслуживания для формирования перечня предлагаемых сервисов; формирование потребительской корзины; принятие решения о выдаче потребительского кредита; сегментирование сферы деятельности с целью повышения эффективности производительности; обработка изображений; тематический анализ библиотеки документов; оптимизация использования складских помещений; выявление транзакций, проведенных по поддельным кредитным картам; выявление потенциальных болезней пациентов; построение показательной (репрезентативной) выборки и т.д.

На сегодняшний день в области кластерного анализа актуально решение следующих проблем:

— обоснованный выбор наиболее подходящего метода исследования, так как он осуществляется из более 100 методов [129];

— сложность оценки получаемых разбиений в целях определения качества проведенного исследования, так как существующие критерии позволяют оценить четкость, компактность, эффективность разбиения, но не решают вопрос выбора оптимального решения для исследуемой предметной области;

— отсутствие рекомендаций по применению существующих методов для использования в исследуемой предметной области – брокерского обслуживания клиентов;

— выбор значения «Количество кластеров», так как данный параметр является входным для большого количества методов.

Прикладной областью диссертационной работы выбрана сфера брокерского обслуживания клиентов. Для обеспечения эффективной работы и конкурентоспособности кредитной организации необходимо регулярное решение целого спектра задач, среди которых можно выделить:

— формирование перечня предлагаемых услуг;

— разработка тарифной политики по взиманию комиссионного вознаграждения за совершение операций в интересах клиентов;

— формирование новых направлений в обслуживании и расширение клиентской базы клиентов.

Значительный рост клиентской базы, на примере одного из крупных банков России, при котором количество клиентов увеличилось в 20 раз, приводит к необходимости использования автоматизированных средств. Однако в настоящее время отсутствуют такие прикладные решения, достаточное количество практических рекомендаций по использованию существующих методов, которые позволяют проводить регулярные исследования интервальной информации об операциях клиентов.

Цель работы и задачи исследования

Целью диссертационной работы является разработка методики адаптивной кластеризации фактографических данных, предназначенной для аналитиков.

Для реализации поставленной цели в работе решаются следующие **задачи**:

1. Исследование методов и систем интеллектуального анализа данных, используемых для кластеризации фактографических данных.
2. Разработка методики адаптивной кластеризации фактографических данных.
3. Разработка рекомендаций по выбору существующих алгоритмов кластеризации.
4. Разработка метода кластеризации.
5. Разработка метода докластеризации.
6. Разработка программного комплекса для автоматизации предложенного метода кластеризации.
7. Оценка эффективности предложенной методики.

Методы исследований

Результаты проведенных и представленных в диссертации исследований получены с использованием теорий классификации и систематизации, алгоритмов, нечетких множеств, графов, реляционных баз данных.

Научная новизна

Научную новизну работы составляют:

— методика адаптивной кластеризации фактографических данных, включающая этап по выбору метода кластеризации;

— метод адаптивной кластеризации (ADAKL) фактографических данных смешанного типа на основе интеграции методов MST и Fuzzy C – Means, позволяющий проводить исследования в выбранной прикладной области, определяя количество и состав кластеров;

— метод докластеризации, позволяющий сократить время кластеризации новых объектов;

— локальный критерий оценки разбиения множества на кластеры, который учитывает требования прикладной предметной области: 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.

Обоснованность и достоверность научных положений, рекомендаций и выводов

Обоснованность научных положений, рекомендаций и выводов определяется корректным использованием математических методов. Достоверность положений и выводов диссертации подтверждается результатами экспериментов.

Практическая ценность

Практическая значимость работы состоит в: формализации методики выполнения кластерного анализа фактографических данных; рекомендациях по использованию существующих и созданного методов кластерного анализа. Практическая ценность разработанного метода состоит в том, что он сокращает время проведения исследования, обладает локальным критерием оценки, позволяющим определить оптимизированное решение задачи. Метод докластеризации позволяет проводить дополнительные исследования новых объектов без проведения общего анализа всех объектов, что приводит к сокращению временных затрат. Кроме того, для исследуемой предметной области – брокерского обслуживания клиентов кредитной организацией также решена задача по выделению существующих групп клиентов, находящихся на обслуживании.

Апробация работы

Основные положения диссертационной работы докладывались и обсуждались на ежегодных заседаниях комиссий по аттестации аспирантов МГТУ им. Н.Э. Баумана в 2006-2008 гг., научных семинарах аспирантов и студентов МГТУ им. Н.Э. Баумана в 2008-2009 г. Апробация работы проводилась на XVI Всероссийской научно – методической конференции «Телематика 2009», на Третьей международной научно – практической конференции «Информационные технологии в образовании, науке и производстве» - 2009, в рамках научной школы «Компьютерная графика и математическое моделирование (Visual Computing)», на семинарах НОК CLAIM, материалы работы представлены для ознакомления и обсуждения с 2008 года на web-сайте и в форуме (электронный адрес - www.philippovich.ru).

Структура и объем работы

Диссертационная работа состоит из введения, четырех глав, заключения, списка использованной литературы и приложений. Общий объем текста диссертации 180 страница, 30 таблиц, 21 схема, 137 источников, в том числе 43 зарубежных.

Содержание работы

Во **введении** описываются основные направления деятельности и специфика решаемых задач кредитной организацией, которые актуализируют задачу по использованию аналитических программных средств в рамках существующих бизнес-процессов. Необходимость в автоматизированных решениях, которые выполняют интеллектуальный анализ данных (ИАД), подтверждается ростом количества систем и их разработчиков. Среди основных задач ИАД выделяются следующие: классификация, регрессия, кластеризация, выявление ассоциаций, выявление последовательностей, прогнозирование. В представляемой диссертационной работе рассматривается задача кластеризации.

Первая глава посвящена исследованию методов кластеризации, аналитических программных комплексов, предметной области.

В работе на основе литературных источников составлена классификация методов кластерного анализа (рис. 1), проведен анализ и сравнение наиболее известных и популярных девяти методов кластерного анализа: CURE, BIRCH, MST, k – средние, PAM, CLOPE, самоорганизующиеся карты Кохонена, HCM, Fuzzy C – Means.

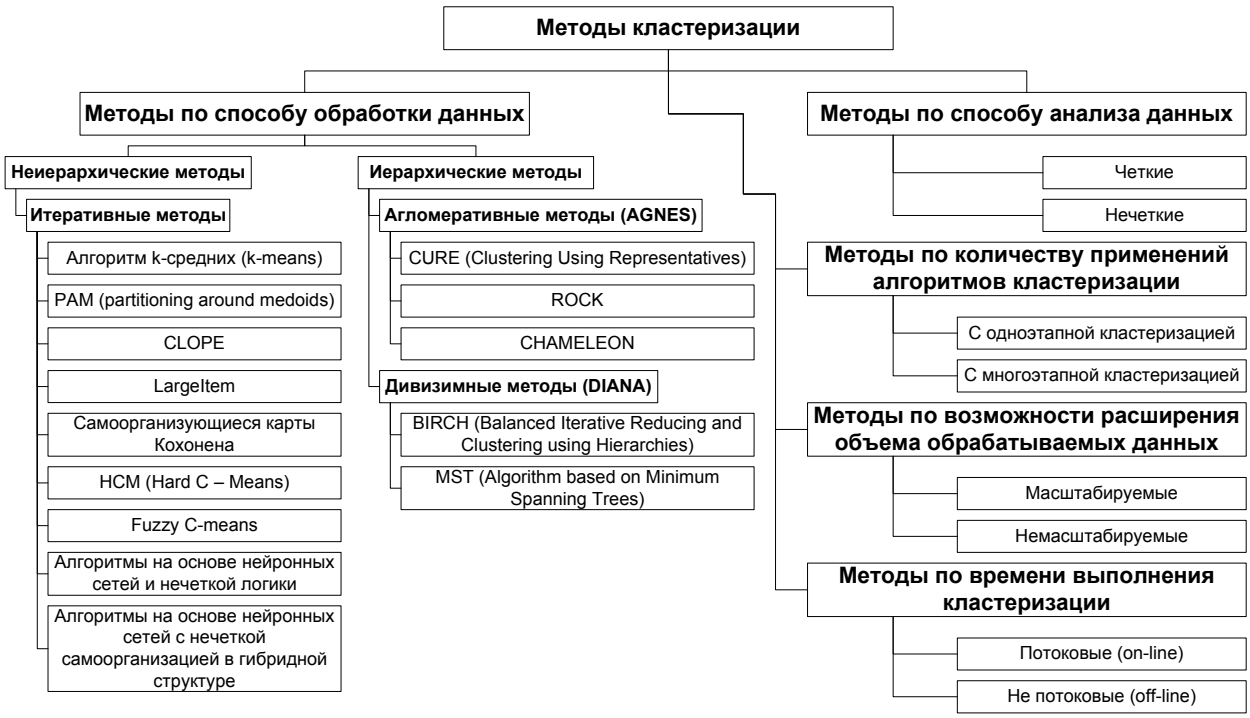

Рис. 1. Классификация методов кластеризации.

Это исследование основано на работах таких авторов как: В. Ганти, И. Герке, Г. Гровэ, С. Гуха, Р. Дюбс, В. Дюк, Б. Дюран, Л. Заде, А. Джэйн, Т. Кормен, Б. Коско, Ч. Лейзерсон, П. Одел, С. Оссовский, К. Парсайе, Р. Рамакришнан, Р. Растоги, Р. Ривест, А. Синг, Ф. Уоссермен, С. Хайкин, Х. Хэ, К. Шим, К. Штайн и др. (всего более 100 источников).

В результате анализа выделены недостатки существующих методов:

— количество кластеров является входным параметром, что приводит к итерационному проведению исследований набора данных и необходимости оценки каждого разбиения на достижение оптимального количества кластеров;

— чувствительность к аномалиям в наборе данных, что требует использования дополнительных инструментов для очищения данных до проведения кластеризации;

— при использовании критериев остановки цикла разбиения на основе разницы между результатами предыдущей и текущей итерации возможны ситуации, при которых происходит зацикливание обработки данных, что приводит к возникновению неопределенностей;

— медленная работа на больших объемах данных, что ограничивает применимость методов;

— нелинейное увеличение времени анализа при росте объемов входных данных, что ведет к значительным временным затратам при динамичном изменении исследуемой сферы деятельности;

— невозможность объяснения полученных результатов разбиения, что снижает доверие у эффективности методов.

На основе проведенного сравнения методов выделены метод MST, который с помощью минимальных остовных деревьев выделяет кластеры произвольной формы, и метод Fuzzy C-means, который выполняет кластеризацию на основе матрицы нечеткого разбиения, что позволяет распределять объекты по одному и более кластерам на основе их степени принадлежности.

При анализе существующих программных комплексов выделены их недостатки:

— в большинстве систем реализовано небольшое количество методов интеллектуального анализа данных, что приводит к необходимости использования большого количества систем и разработки интеграционных решений для сравнения эффективности методов;

— большинство аналитических программных комплексов не имеют комплексного подхода при обработке данных, из-за отсутствия которого появляется необходимость в использовании других инструментов для подготовки входных данных, что ведет к увеличению временных затрат на исследование.

Предметной областью для исследования, в которой решается практическая задача, является брокерское обслуживание клиентов. На основе проведенного исследования выделен и формализован класс задач:

— Количество исходных объектов: $K = \left[Клиент_1, Клиент_2, ..., Клиент_i \right]$, $i \in \left[500; 50000 \right]$;

— Количество значимых характеристик объектов: $Клиент_i = \left[k_1, k_2, ..., k_j \right], j \in \left[70; 150 \right]$;

— Типы характеристик T : $T \in \left[числовые, лингвистические \right]$;

— Форма получаемых кластеров – сложная, с пересечениями;

— Количество кластеров N – результат исследования: $N \in \left[5; 30 \right]$.

В результате анализа среди существующих методов кластеризации выявлены методы адаптивной кластеризации, под которыми в работе понимаются методы, входной параметр «Количество кластеров» которых является результатом исследования за счет оптимизации локальных критериев оценки качества разбиения.

Вторая глава посвящена постановке задачи кластеризации, построению формализованной модели предметной области, исследованию и адаптации существующих методов кластеризации фактографических данных.

Формализованную модель предметной области можно спроецировать в реляционную структуру взаимоотношений с применением логического преобразования и обобщения. Полученная в результате модель состоит из двух основных сущностей: клиент, операции клиента с различными финансовыми инструментами, т.е.:

$$\mathbf{Cl_{Type}} = \gamma\left(\left\{ \left\langle \mathrm{Cl_{PersInfo}}, \mathrm{Cl_{ListUslug}} \right\rangle \right\} \right), \ \gamma \to \mathrm{Optimum}, \text{ где}$$

$\mathbf{Cl_{Type}}$ - перечень обслуживаемых групп клиентов;

$\mathrm{Cl_{PersInfo}}$ - перечень предоставляемой персональной информации;

$\mathrm{Cl_{ListUslug}}$ - перечень предоставляемых услуг в рамках брокерского обслуживания.

В зависимости от характеристик исходных данных и желаемого результата на данный момент существуют следующие инструменты для проведения исследований: статистика, деревья решений, нейронные сети, теория графов, нечеткая логика.

Статистические методы, как и деревья решений, теория графов, применяются в задачах, когда данные хорошо разделимы и не содержат большого количества выбросов. Такие методы, как нейронные сети и нечеткая логика, позволяют работать с большими наборами данных, которые имеют выбросы. Методы, основанные на нечеткой логике, позволяют делать мягкое разделение массива данных на кластеры, определяя одни и те же данные в разные кластеры с разной степенью принадлежности. Эта особенность нечетких методов позволяет применять методы на неоднородных данных и получать при этом довольно качественные результаты.

Нейронные сети и нечеткая логика являются лидерами среди методов анализа данных большого объема со значительным количеством анализируемых атрибутов.

В связи с тем, что на данный момент существующих практических рекомендаций по использованию методов недостаточно и количество их достаточно велико, то была разработана методика адаптивной кластеризации фактографических данных (рис. 2), которая направлена на решение этой задачи. Она описывает процесс проведения анализа массива фактографических данных от первого этапа – выборки исходных данных для проведения анализа, до последнего этапа – получения итоговых результатов в виде разбиения на кластеры.

Рис. 2. Методика адаптивной кластеризации.

Выборка исходных данных может производиться с помощью различных средств: путем построения регулярных запросов, ведения сведений в различных системах оперативного, аналитического учета и т.п. Полученная выборка на втором этапе подлежит исследованию с целью выявления значимых объектов/характеристик объектов, которое выполняется с помощью существующих методов, например, понижения размерности с помощью факторного анализа, устранения незначащих характеристик с помощью корреляционного анализа, выявления дубликатов и противоречий и т.п. Данный этап позволяет сократить временные затраты на выполнение исследования за счет уменьшения объемов исследуемого массива информации, а также повысить эффективность исследования за счет исключения из выборки противоречивых данных. На основе полученных данных есть возможность разработать контрольный пример, который в дальнейшем будет использован для проверки действенности метода. Данный процесс необходимо выполнять с привлечением носителей экспертных знаний в исследуемой области. На следующем этапе выполняется выбор метода кластерного анализа. При выборе метода проведения исследования есть возможность использовать существующие методы кластеризации или использовать авторский метод адаптивной кластеризации ADAKL. Характерной особенностью данного этапа является то, что на основе характеристик полученной выборки и априорным знанием желаемого результата

выполняется поиск подходящего метода исследования с промежуточными оценками результатов и накоплением практического опыта по применению различных методов в решаемой практической задаче. После выбора метода кластерного анализа выполняется кластеризация полного объема данных и получение результата в виде конечного разбиения множества исходных объектов на кластеры.

Выбор существующего метода кластеризации выполняется тремя шагами: выбор метода, настройка параметров выбранного метода, анализ массива исходных данных и оценка результатов исследования. Выбор метода может быть осуществлен тремя способами: на основе существующих рекомендаций, полученных в результате анализа литературных источников, на основе критериев, по общему алгоритму за счет перебора существующих методов. На основе литературных источников выделено восемь критериев выбора: объем информации по отношению к времени обработки (Cr_1), размерность информации (Cr_2), типы атрибутов сущностей (Cr_3), чувствительность к равномерности информации (Cr_4), априорное (экспертное) представление о форме получаемых кластеров (Cr_5), априорное (экспертное) представление о количестве кластеров (Cr_6), априорное (экспертное) представление о перекрываемости кластеров (Cr_8). С точки зрения исследуемой предметной области критерии в соответствии с требованиями решаемой задачи следующие: $Cr_8 = Высокое$, $Cr_6 = Вычисляемая\ величина$, $Cr_5 = Сложная\ форма$, $Cr_7 = С\ пересечениями$, $Cr_2 = Высокая\ размерность$, $Cr_3 = Смешанного\ типа$, $Cr_1 = Линейная\ или\ логарифмическая\ зависимость$, $Cr_4 = Низкая\ чувствительность$.

Входные параметры, с помощью которых выполняется настройка методов для использования в предметной области, можно разделить на характеристические, итерационные и экспертные. Сложность процесса настройки алгоритма заключается в том, что требуется решить задачу, которая носит итерационный характер, нахождения приемлемого баланса между характеристиками предметной области и возможностями настройки алгоритма кластеризации.

При анализе источников литературы выявлены следующие рекомендации по использованию методов для исследования предметных областей и практических задач:

Табл. 1. Выбор метода кластеризации на основе рекомендаций из источников.

Метод	Рекомендации к применению	Противопоказания к применению
CURE	Выявляет кластеры произвольной формы, метод менее чувствителен к выбросам, чем MST. Время работы алгоритма незначительное.	
BIRCH	Метод предназначен для очень больших наборов данных. Работает с произвольным количеством оперативной памяти. Получаемое разбиение обладает высоким качеством.	Не использовать для получения несферических или широковарьируемых форм кластеров.
MST	Лучше всего подходит для произвольных форм.	Очень чувствителен к выбросам.
k – средних	Хорошо работает с хорошо разделимыми данными и выделяет кластеры сферической формы.	
PAM	Хорошо работает с хорошо разделимыми данными и выделяет кластеры сферической формы.	
CLOPE	Кластеризация огромных объемов категорийных данных.	
Самоорганизующиеся карты Кохонена	Поиск и анализ закономерностей.	
HCM	Хорошо работает с хорошо разделимыми данными и выделяет кластеры сферической формы.	
Fuzzy C – Means	Относит объект к разным кластерам на основе степени принадлежности элемента к кластерам, выделяет кластеры сферической формы.	

Сложностью кластеризации является необходимость оценки её результатов, которая позволяет определить возможность использования алгоритма для выбранной предметной области. Оценка качества кластеризации проверяет результаты анализа в качественной и объектной формах. В рамках исследования может выполняться два вида оценки: экспертная и формальная. Экспертная оценка включает ручную проверку, усреднение характеристик объектов и оценку их удаленности, проверку результатов на контрольном примере, добавление новых объектов и оценку стабильности разбиения, использование различных методов и сравнение результатов разбиения. Формальная оценка выполняется

на основе формализованных критериев, например, индекса «Хие-Бени», индекса истинности разбиения, коэффициента разбиения, индекса четкости, показателя компактности и изолированности, индекса эффективности.

Аналитическая оценка сложности метода MST зависит от используемого алгоритма построения минимального остовного дерева: по алгоритму Борувки - $O\left[E \cdot Log\left(V\right)\right]$, по алгоритму Крускала - $O\left[E \cdot Log\left(E\right)\right]$, по алгоритму Прима - $O\left[E \cdot Log\left(V\right)\right]$, где V - множество вершин графа, E - множество их возможных попарных соединений (ребер). Аналитическая оценка метода Fuzzy C-means имеет линейную зависимость относительно количества кластеров и исследуемых объектов, но сложность этой оценки в том, что метод представляет собой циклическую структуру с условием выхода из цикла по параметру останова.

Третья глава посвящена разработке метода адаптивной кластеризации фактографических данных смешанного типа ADAKL на основе дивизимных и итерационных методов. За основу данного метода в части первичного разделения объектов на кластеры взят метод MST, использующий минимальные остовные деревья, и метод Fuzzy C-means. Определяющими факторами в выбранной комбинации является способность при использовании теории графов выделять кластеры произвольной формы и оптимальной структуры, а при использовании математического аппарата нечеткой логики решается задача разделения объектов с лингвистическими атрибутами. Также в методе используются математические разработки ученых в части решения задачи нахождения минимального остовного дерева [95].

Совокупность использованных методов и алгоритмов позволяет преодолеть недостатки каждого из них: для MST – применение нечеткости позволяет сделать более плавное разбиение, помещая объекты в разные кластеры с разной степенью принадлежности, для Fuzzy C–Means – предварительное использование MST и модифицированного критерия оптимальности позволяет сократить количество итераций исследования входного набора данных, а следовательно, и снизить временные, человеческие и технические затраты на проведение исследований.

При работе ADAKL строится минимальное остовное дерево, образуя оптимизированную древовидную структуру из исходных элементов на основе характеристик кластеризуемых объектов, и выделяются первичные кластерные центры. Затем используется итерационный подход, с помощью которого уточняются центры кластеров и содержимое кластеров на основе вычисления степени принадлежности объекта кластеру. ADAKL состоит из пяти этапов: нормализация числовых атрибутов, вычисление матрицы взаимных расстояний между объектами, построение минимального остовного дерева, разделение объектов на кластеры и построение матрицы нечеткого разбиения, выбор наилучшего разбиения. ADAKL использует скрытые зависимости между объектами входного набора данных и позволяет решать задачу кластерного анализа объектов с атрибутами смешанного типа с использованием предварительно настроенной словарной системы и нечеткой логики при определении соотношений между понятиями. Двухэтапность выполнения кластеризации и использование модифицированного критерия оптимальности позволяет повысить качество проводимой кластеризации.

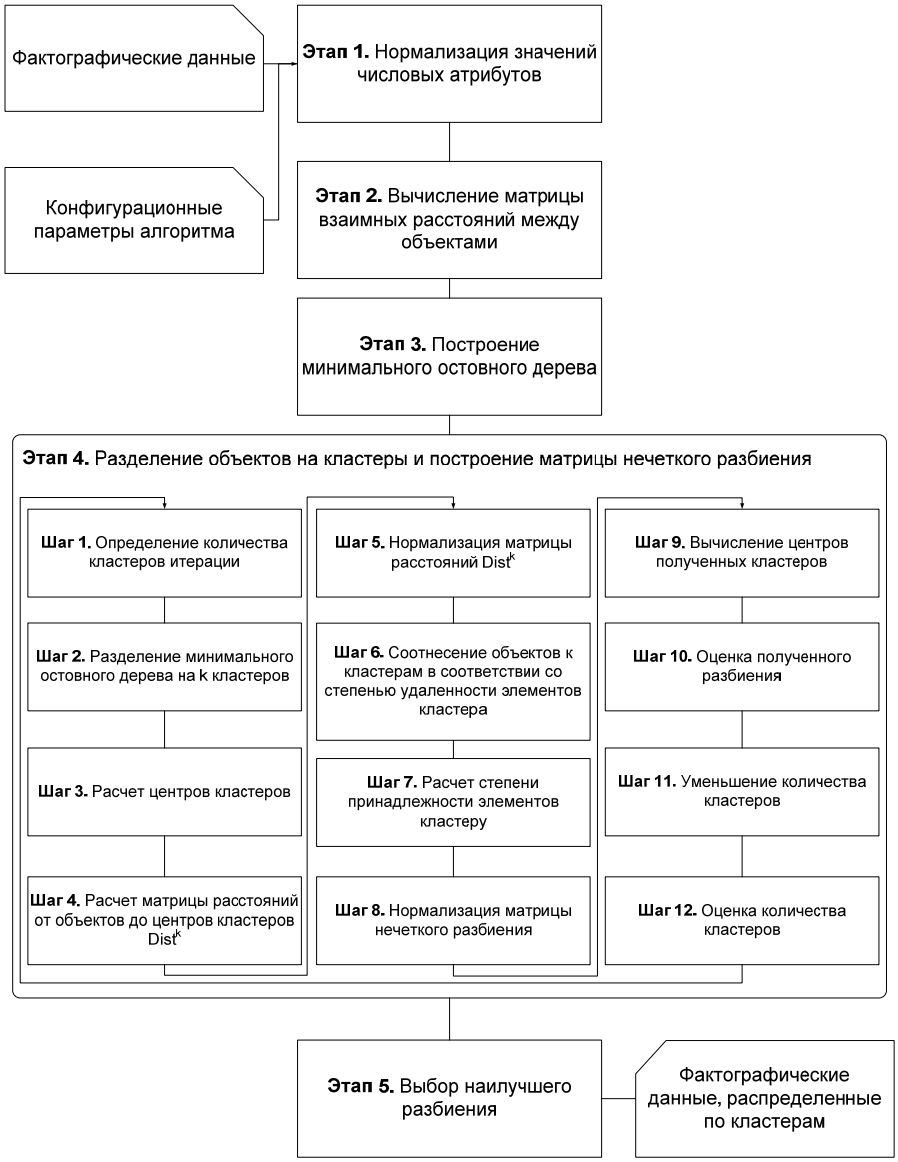

Рис. 2. Основные этапы метода адаптивной кластеризации ADAKL.

Для устранения чувствительности к выбросам на первом этапе предлагается использование предобработки исходных данных через нормализацию:

— линейная нормализация:

$$Value_{ij} := \left\{ \frac{Value_{ij}}{\underset{j}{Max}\left(Value_{ij}\right)} \middle| \begin{matrix} \underset{j}{Max}\left(Value_{ij}\right) \neq 0, \\ t_j \in \{Целочисленный\ тип, Денежный\ тип\} \end{matrix} \right\}$$

— статистическая нормализация:

$$Value_{ij} := \left\{ \frac{\left| Value_{ij} - \frac{\sum_{i=1}^{m} Value_{ij}}{m} \right|}{\sqrt{\frac{\sum_{i=1}^{m}\left(Value_{ij}\right)^2}{m^2} - \left(\frac{\sum_{i=1}^{m} Value_{ij}}{m}\right)^2}} \middle| \begin{matrix} \sqrt{\frac{\sum_{i=1}^{m}\left(Value_{ij}\right)^2}{m^2} - \left(\frac{\sum_{i=1}^{m} Value_{ij}}{m}\right)^2} \neq 0, \\ t_j \in \begin{Bmatrix} Целочисленный\ тип, \\ Денежный\ тип \end{Bmatrix} \end{matrix} \right\}$$

При вычислении информационных расстояний между объектами используются классические формулы вычисления расстояний, доработанные для использования в методе:

$Dist_{ij} = \left\| u_i - u_j \right\| = Metric\left(u_i, u_j\right)$, где $Metric$ – способ определения расстояния между объектами.

Если $Metric = Евклидово\ расстояние$, то $Dist_{ij} = \sqrt{\sum_w \left(\left[Value_{iw} - Value_{jw}\right] * K_w\right)^2}$,

Если $Metric = Квадрат\,Евклидова\ расстояния$, то $Dist_{ij} = \sum_w \left(\left[Value_{iw} - Value_{jw}\right] * K_w\right)^2$,

Если $Metric = РасстояниеЧебышева$, то $Dist_{ij} = \underset{w}{Max}\left[Value_{iw} - Value_{jw}\right] * K_w$,где $i, j \in \left[1, m\right], w = \overline{1,n}$ при условии $FieldType\left[w\right] = "Входное"$

Построение минимального остовного дерева на третьем этапе может быть выполнено тремя способами: алгоритм Борувки $\left[O\left(\left\|Dist\right\|^2 \cdot Lg\left(m\right)\right)\right]$, алгоритм Крускала $\left[O\left(\left\|Dist\right\|^2 \cdot Lg\left\|Dist\right\|\right)\right]$, алгоритм Прима $\left[O\left(\left\|Dist\right\|^2 \cdot Lg\left(m\right)\right)\right]$. Результат работы приведенных алгоритмов одинаков, но они обладают разной вычислительной сложностью, что является критичным при анализе значительных объемов данных, поэтому для оптимизации временных затрат на исследование необходимо применять алгоритм Прима.

На основе построенной оптимизированной структуры объектов в виде дерева строится матрица нечеткого разбиения, которая обладает следующими характеристиками:

$F = \left[\mu_{ij}\right], \mu_{ij} \in \left[0,1\right], i \le q, j = \overline{1,m}$, где μ_{ij} – степень принадлежности $i^{го}$ объекта к $j^{му}$ кластеру. Матрица разбиения обладает следующими свойствами: $\sum_{i=1}^{k} \mu_{ij} = 1, j = \overline{1,m}$, $0 < \sum_{j=1}^{m} \mu_{ij} \le m, i = \overline{1,k}$.

На третьем шаге четвертого этапа выполняется первичное выделение центров кластеров с помощью следующего выражения:

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

где Avg – оператор вычисления среднего значения показателей объектов, входящих в кластер k, $i = \overline{1,k}$, $j = \overline{1,m}$.

На следующем шаге выполняется расчет матрицы расстояний от объектов до центров кластеров V_i^k:

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

где $i = \overline{1,k}$, $j = \overline{1,m}$, $Metric$ – способ определения расстояния между объектами.

Нормализация матрицы расстояний от объектов до центров кластеров V_i^k выполняется на основе формулы:

$$Dist_{ij}^{k'} = \begin{cases} \dfrac{Dist_{ij}^k}{Max\left(Dist_{ij}^k\right)}, Max\left(Dist_{ij}^k\right) \ne 0 \\ Dist_{ij}^k, Max\left(Dist_{ij}^k\right) = 0 \end{cases}$$, где $i = \overline{1,k}$, $j = \overline{1,m}$.

При соотнесении объектов к кластерам в соответствии со степенью удаленности элементов кластера (w) используется следующее выражение:

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

После разнесения объектов по кластерам выполняется расчет степеней принадлежности к кластерам текущей итерации алгоритма:

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

По итогам завершения распределения объектов выполняется нормализация полученной матрицы нечеткого разбиения:

$$\mu_{ij} = \left\{ \begin{array}{l} \dfrac{\mu_{ij}}{\sum\limits_{i=1}^{k} \mu_{ij}} \left| \sum\limits_{i=1}^{k} \mu_{ij} \ne 0 \right. \\ \mu_{ij} \left| \sum\limits_{i=1}^{k} \mu_{ij} = 0 \right. \end{array} \right\}, \text{ где } j = \overline{1,m}.$$

На основе полученной матрицы нечеткого разбиения выполняется вычисление новых центров кластеров с учетом последнего перераспределения объектов:

$$V_i^{k'} = \frac{\sum\limits_{\substack{j=1 \\ u_j \in V_i^{k'}}}^{m} \mu_{ij}^{\ p} * u_j}{\sum\limits_{\substack{j=1 \\ u_j \in V_i^{k'}}}^{m} \mu_{ij}^{\ p}}, \text{где } i = \overline{1,k}.$$

На следующем шаге оценивается качество полученного разбиения на k кластеров с использованием полученных центров:

$$O^k = \frac{\sum\limits_{i=1,k} \dfrac{\left|V_i^{k'}\right| * \sum\limits_{\substack{j=1 \\ u_j \in V_i^{k'}}}^{m} \mu^p{}_{ij} * \left\|V_i^{k'} - u_j\right\|}{\underset{\substack{i \ne j \\ u_j \in V_i^{k'}}}{Min}\left(\left\|V_i^{k'} - u_j\right\|\right) * \underset{u_j \in V_i^{k'}}{Max}\left(\left\|V_i^{k'} - u_j\right\|\right) * \sum\limits_{\substack{j=1 \\ u_j \in V_i^{k'}}}^{m} \left\|V_i^{k'} - u_j\right\| * k}}{m * k^2}, \text{ где}$$

$\left|V_i^{k'}\right|$ – количество элементов в кластере i;

$\left\|V_i^{k'} - u_j\right\| = Metric\left(V_i^{k'}, u_j\right)$ – расстояние от центра кластера i до элемента u_j;

$u_j \in V_i^{k'}$ – отражение условия о принадлежности элемента кластеру.

Предложенная оценка является составной:

$$O^k = \frac{\sum\limits_{i=1,k} \dfrac{\boxed{1}\ \left|V_i^{k'}\right| * \boxed{3}\ \sum\limits_{\substack{j=1 \\ u_j \in V_i^{k'}}}^{m} \mu^p{}_{ij} * \left\|V_i^{k'} - u_j\right\|}{\underset{\substack{i \ne j \\ u_j \in V_i^{k'}}}{Min}\left(\left\|V_i^{k'} - u_j\right\|\right) * \underset{u_j \in V_i^{k'}}{Max}\left(\left\|V_i^{k'} - u_j\right\|\right) * \sum\limits_{\substack{j=1 \\ u_j \in V_i^{k'}}}^{m} \left\|V_i^{k'} - u_j\right\| * k}}{\boxed{2}\ m * k^2}$$

Область 1 – нацелена на выделение кластеров с наименьшими взаимными расстояниями и наибольшим количеством элементов в кластере по отношению к общему количеству кластеров.

Область 2 – выделяет количество получаемых кластеров и ведет к уменьшению их количества.

Область 3 – нацелена на минимизацию взаимных расстояний между полученным центром кластера и элементами с учетом степени принадлежности.

Выбор наилучшего разбиения по результатам всех итераций выполняется на основе лучшей оценки: $O_{Опт} = \underset{i=\overline{1,q}}{MAX}\left(O^i\right)$, где $i = \overline{1,q}$.

Предложенный метод обладает квадратичной зависимостью аналитической сложности алгоритма от количества исходных данных по объектам кластеризации, что существенно увеличивает временные затраты при регулярном появлении новых данных и повторной кластеризации.

Частично преодолеть этот недостаток можно за счет специальной процедуры докластеризации, которая определяет необходимость повторного запуска исследования полного массива данных и, в случае отсутствия признаков появления новых значимых групп объектов, осуществляет распределение новых (расширяющих) объектов по имеющимся кластерам. Для расширения исходных данных в процессе проведения анализа необходимо произвести дополнительное исследование добавляемых данных (рис. 3). Процесс докластеризации интегрируется с основным алгоритмом, получая возможность выполнять некоторые этапы независимо от основного алгоритма.

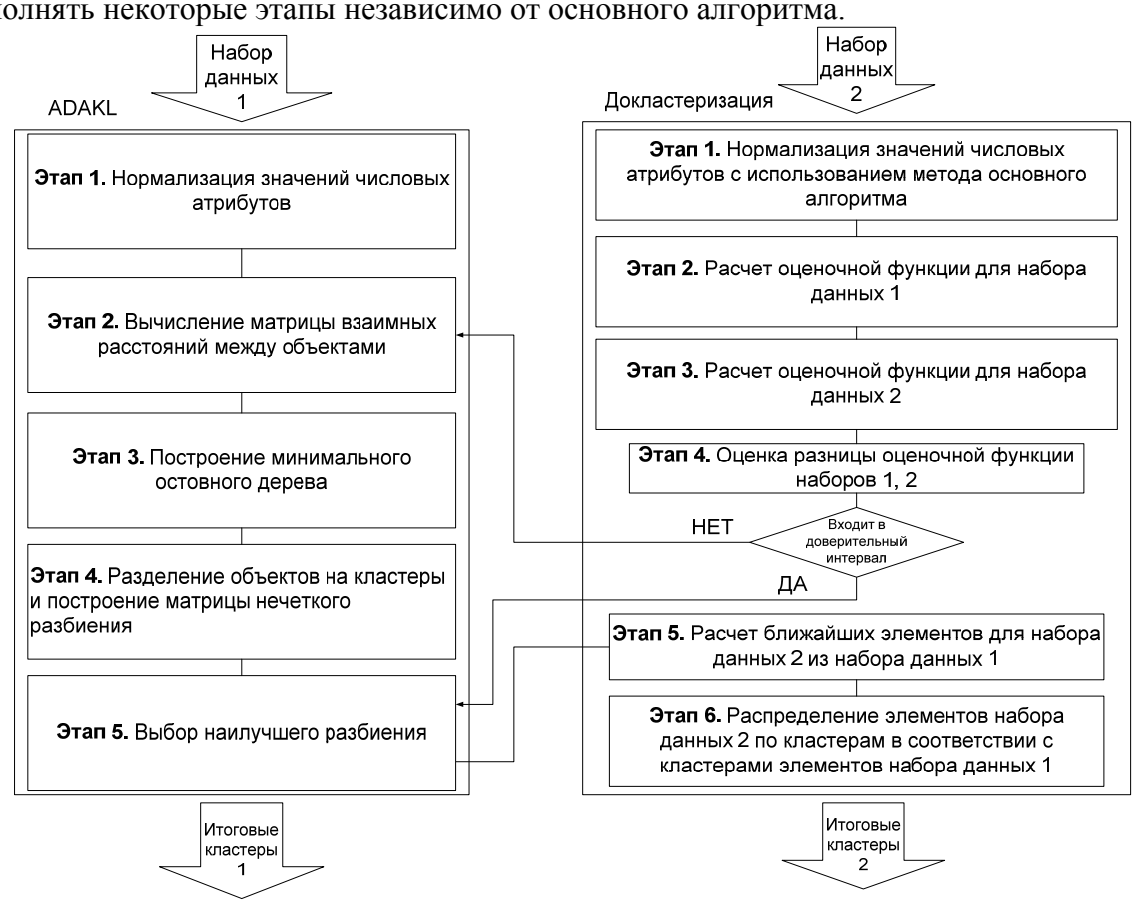

Рис. 3. Докластеризация исходных данных.

Необходимость в докластеризации подтверждается результатами эмпирических исследований, по результатам которых выявлено, что наиболее трудоемким этапом метода является построение минимального остовного дерева (рис. 4). Выполнение дополнительного исследования при расширении исходных данных позволяет значительно сократить временные затраты по анализу данных за счет распределения расширяющих объектов по имеющимся кластерам в случае подобности исходных данных в наборах 1, 2.

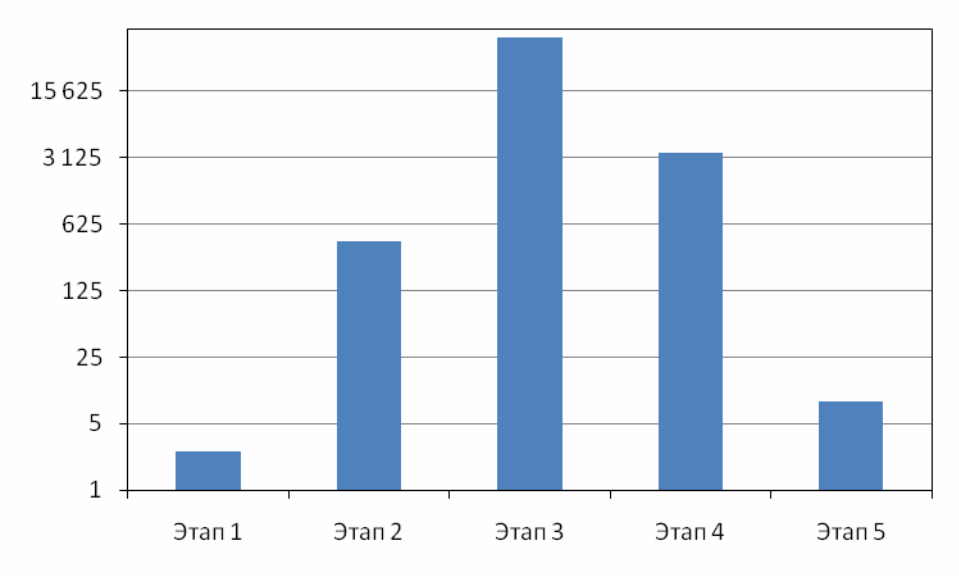

Рис. 4. Общее время выполнения анализа в разрезе этапов.

Принятие решения о возможности распределения объектов по полученным кластерам в результате основного исследования выполняется на основе разницы оценочных функций:

$$O_1 = \sqrt{\frac{\sum_{i=1}^{r} \left\| A_i - Avg[A] \right\|^2}{r}}, \; O_2 = \sqrt{\frac{\sum_{i=1}^{o} \left\| B_i - Avg[B] \right\|^2}{o}}, \text{ где}$$

O_1, O_2 – оценочная функция исходного набора данных 1 и 2 соответственно;

A, B – исходные наборы данных 1 и 2 соответственно;

r, o – количество объектов в исходных наборах данных 1 и 2 соответственно;

$\left\| A_i - Avg[A] \right\|, \left\| B_i - Avg[B] \right\|$ – оператор вычисления расстояния между объектом и средним значением множества, полученного с использованием оператора вычисления среднего значения основного алгоритма, для исходных наборов данных 1 и 2 соответственно. Вычисление данного расстояния выполняется с учетом весовых коэффициентов основного алгоритма: $K = \{K_1, K_2, ..., K_n\}$, где K_i – весовой коэффициент влияния атрибута объекта, $K_i \in [0;1]$.

Пороговое значение, обозначающее подобность обоих наборов данных, является входным параметром ADAKL.

Распределение элементов расширяющего множества (B) по вычисленному расстоянию до ближайших k – объектов[1] из расширяемого множества (A) выполняется следующим образом:

$\mu_{ij} = \left(1 - Dist_k^{Norm}\right) * \mu_{kj} = \left[1 - Dist_k^j / Max\left(Dist_k\right)\right] * \mu_{kj}$, где

$i = \overline{1, o}$ – порядковый номер элемента из множества B;

$j \in [1, r]$ – порядковый номер ближайшего элемента из множества A для соответствующего элемента из множества B;

k – порядковый номер ближайшего элемента;

$Dist_k^j = \left\| B_i - A_j \right\|$ – расстояние между ближайшими элементами из множеств A и B;

[1] k – оптимальное количество кластеров в соответствии с критерием: $O_{Onm} = \underset{i=1,q}{MAX}\left(O^i\right)$

$Max\left(Dist_k\right)$ – максимальное расстояние от элемента из множества B до элемента из множества A;

μ_{kj} – степень принадлежности элемента из множества A к кластеру k.

В результате анализа аналитической сложности метода получены следующие оценки: с линейной нормализацией – $O\left(m^2 * \left(a+b+Lg\left(m\right)+q\right)\right)$, со статистической нормализацией – $O\left(m^2 * \left(a+b+Lg\left(m^2\right)+q\right)\right)$, и докластеризации: $O\left(\left(m+n\right)^2 * \left[a+b\right]\right)$, где a – количество входных числовых атрибутов, b – количество входных лингвистических атрибутов, m – количество кластеризуемых объектов набора данных 1, n – количество кластеризуемых объектов набора данных 2, q – общее количество кластеров. Основным недостатком разработанного метода является квадратичная зависимость аналитической сложности от количества исходных данных по объектам кластеризации.

Предложенный метод обладает следующими достоинствами:

— двухэтапная кластеризация, которая позволяет выделить большее количество закономерностей;

— способен работать с лингвистическими атрибутами объектов, позволяя решить проблему использования экспертных оценок и текстовых атрибутов объектов;

— использует весовые коэффициенты для анализируемых атрибутов, позволяя не менять результирующий набор данных и работать со всем массивом, варьируя влиянием атрибута на результат анализа;

— использует степень удаленности объектов/элементов, позволяя соотносить объекты по кластерам при разделении на основе вычисленного расстояния;

— использует размазанность кластера, которая позволяет определять четкость получаемых границ кластеров;

— использует критерий оценки разбиения на кластеры, который учитывает требования и специфику предметной области;

— способен выполнить докластеризацию дополнительного набора данных, позволяя сократить временные затраты на анализ данных в случае необходимости добавления объектов к основному массиву данных за счет дополнительного исследования только расширяющих объектов вместо перезапуска всего исследования.

Четвертая глава посвящена описанию программного решения (ПР), реализующего ADAKL.

Целью реализации метода в виде ПР является автоматизация процесса обработки исходных данных по разработанным алгоритмам для проведения практических исследований в процессе опытной эксплуатации, а затем и для промышленного использования данного программного обеспечения в компании. Данное ПР можно отнести к типу интегрированных решений ввиду его функциональных возможностей. Также в этой главе описывается архитектура ПР и основные алгоритмы, реализованные в ПР. Инфологическая/даталогическая модели ПР, приведенные в приложении 5, предусматривают хранение основных сущностей, необходимых для настройки и запуска анализа, а также сущности, которые позволяют сохранить результаты исследования массивов для последующего сравнительного анализа. ПР позволяет сохранять промежуточные, итоговые результаты анализа и используемые данные в следующих форматах: текстовый, гипертекстовой разметки, MS Excel 2003.

В пятом разделе главы описываются три основных и одна дополнительная серии по пятьдесят экспериментальных исследований для оценки работоспособности ADAKL в сравнении с другими алгоритмами: 1) выделение секторов инвестирования на основе анализа показателей финансовых инструментов; 2) выделение групп клиентов на основе статистических данных о деятельности клиентов за период; 3) выявление категорий финансовых инструментов для оценки эффективности операций; 4) выделение классов

автомобилей на основе данных о максимальной скорости, цвете кузова, воздушном сопротивлении, массе. Исследование производилось на трех методах: 1 – самоорганизующиеся карты Кохонена, 2 – алгоритм k – средних, 3 – разработанный метод. Исследуемые массивы данных имеют следующие характеристики:

Табл. 2. Сводная таблица исследуемых 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]

Оценка разбиения выполнена на основе показателей выполненной кластеризации с помощью индекса истинности разбиения:

$$O = \frac{r}{n} * \begin{cases} q/k, q \le k \\ k/q, q > k \end{cases}, \text{ где}$$

q – количество кластеров по итогам кластеризации;

r – количество элементов, правильно распределенных по соответствующим кластерам;

k – исходное количество кластеров;

n – количество объектов кластеризации.

По результатам исследования составлена сводная таблица с усредненными оценками разбиений алгоритмами:

Табл. 3. Средневзвешенная оценка разбиений.

Оценка / Метод	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
1	0.7913	0.9150	0.9237	0.8767
2	-	0.8232	-	0.8232
3	0.9762	0.9981	0.9990	0.9911

В соответствии с полученной итоговой оценкой наилучшее разбиение на исследованных массивах по сериям экспериментов получено с применением разработанного метода ADAKL. Проведенные эксперименты подтвердили, что использование интеграции методов кластеризации (многоэтапная кластеризация) улучшает качество выявления знаний в сравнении с одноэтапными методами, а также то, что превосходство разработанного метода достигается использованием математического аппарата нечеткой логики и внутренних словарей системы при определении информационных расстояний между объектами.

Для проверки аналитической оценки метода проведено пять серий по пятьдесят нагрузочных экспериментов на основе эмпирических данных. Нелинейность и её порядок относительно количества записей входного набора данных, полученные в результате четырех серий экспериментов и представленные на рис. 5, 6, подтверждают полученную ранее аналитическую оценку алгоритма.

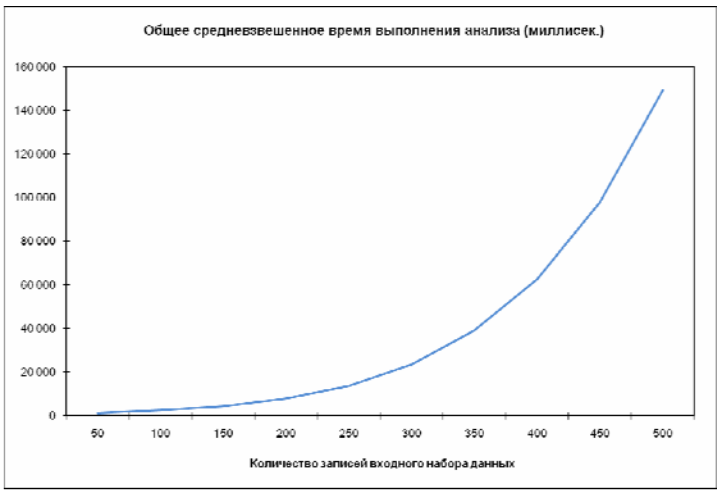

Рис. 5. Общее средневзвешенное время анализа массива.

Рис. 6. Общее средневзвешенное время докластеризации массива.

На основе метода ADAKL разработано ПР, с помощью которого выполнено выделение групп клиентов и определение их доли от общего количества клиентов (рис. 7).

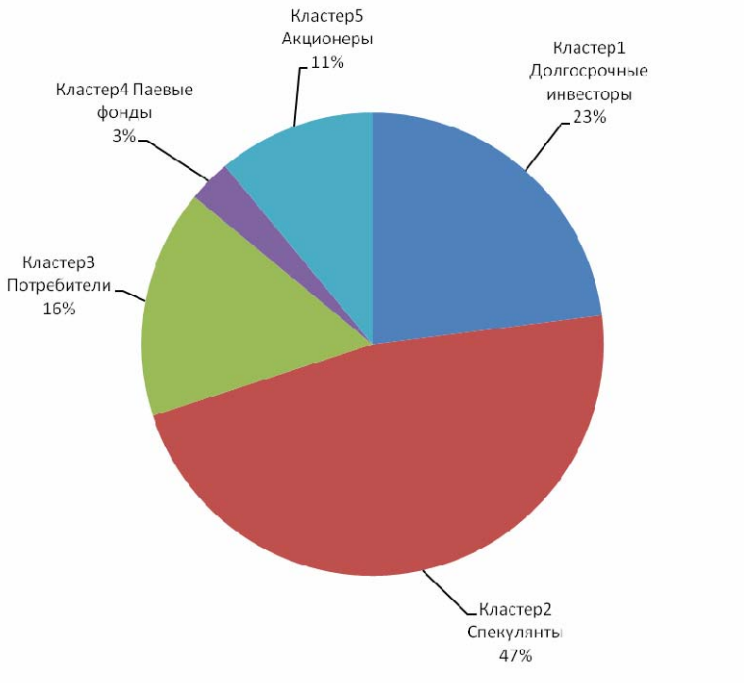

Рис. 7. Распределение клиентов по группам.

Последующий анализ экономических показателей полученных групп объектов позволил дать названия кластерам: Кластер 1 «Долгосрочные инвесторы» - 23%, Кластер 2

«Спекулянты» - 47%, Кластер 3 «Потребители» - 16%, Кластер 4 «Паевые фонды» - 3%, Кластер 5 «Акционеры» - 11%, и разработать более целевую, направленную на конкретную клиентскую группу тарифную политику, а также предложить им более выгодные условия по совершаемым видам операций, увеличив количество этих операций и объем комиссионных сборов, что положительно повлияет на доходность данного направления деятельности кредитной организации.

Основные выводы и результаты диссертационной работы

Совокупность сформулированных и обоснованных в диссертации методов и положений, а также её практические результаты представляют собой решение актуальной научно-технической задачи извлечения закономерностей из фактографических данных смешанного типа. Сформулированные положения и разработанный метод адаптивной кластеризации позволяют автоматизировать процесс выбора метода выполнения кластерного анализа данных в выбранной предметной области, а также повысить эффективность и качество кластеризации за счет интеграции методов кластерного анализа.

Основные результаты диссертационной 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 MST и Fuzzy C – Means, 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).

Публикации

Статьи, опубликованные в ведущих рецензируемых научных журналах и изданиях, определенных ВАК

1. Нейский, И.М., Филиппович, А.Ю. Методика адаптивной кластеризации фактографических данных на основе интеграции алгоритмов MST и Fuzzy C-means / И.М. Нейский, А.Ю. Филиппович // Известия высших учебных заведений. Проблемы полиграфии и издательского дела. – М.: Изд-во МГУП, 2009. – №3 – С. 48-61.

Другие публикации

2. Нейский, И.М. Характеристика технологий и процессов интеллектуального анализа данных / И.М. Нейский // Интеллектуальные технологии и системы. Сборник учебно-методических работ и статей аспирантов и студентов. – М.: Изд-во ООО «Эликс+», 2006. – Выпуск 7. – С. 111-122.

3. Нейский, И.М. Классификация и сравнение методов кластеризации / И.М. Нейский // Интеллектуальные технологии и системы. Сборник учебно-методических работ и статей аспирантов и студентов. – М.: Изд-во ООО “Эликс +”, 2008. – Выпуск 8. – С. 111-122.

4. Нейский, И.М., Филиппович, А.Ю. Интеграция дивизимных и итерационных методов для адаптивной кластеризации фактографических данных / И.М. Нейский, А.Ю. Филиппович // Труды конференции «Телематика`2009» – М.: 2009. – С. 413-414.

5. Нейский И.М. Адаптивная кластеризация на основе дивизимных и итерационных методов / И.М. Нейский // Сборник трудов третьей международной научно – практической конференции «Информационные технологии в образовании, науке и производстве» под редакцией Ю.А. Романенко. – МО., [2009]. – С. 172-175.

6. Нейский И.М. Докластеризация как способ оптимизации времени анализа исходных данных / И.М. Нейский // Научная школа для молодых ученых «Компьютерная графика и математическое моделирование (Visual Computing)»: тезисы и доклады. – М., 2009. – С. 141-161.