На правах рукописи

Нейский Иван Михайлович

МЕТОДИКА АДАПТИВНОЙ КЛАСТЕРИЗАЦИИ ФАКТОГРАФИЧЕСКИХ ДАННЫХ НА ОСНОВЕ ИНТЕГРАЦИИ МЕТОДОВ МИНИМАЛЬНОГО ОСТОВНОГО ДЕРЕВА И НЕЧЕТКИХ К-СРЕДНИХ

05.13.17 – Теоретические основы информатики

АВТОРЕФЕРАТ
диссертации на соискание ученой степени
кандидата технических наук

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: доктор технических наук, профессор
Ковшов Евгений Евгеньевич
кандидат технических наук,
Паклин Николай Борисович

Ведущая организация: ГОУ ВПО «Петрозаводский государственный университет»

Защита диссертации состоится 25 ноября 2010 г. в 13:00 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 «_____» октября 2010 г.

Ученый секретарь
диссертационного совета
д.т.н., профессор

В.Н. Агеев

Общая характеристика работы

Актуальность работы

В настоящее время, в связи с широким применением в обществе информационных технологий на базе использования средств вычислительной техники, сформировалась область научных исследований и задач разработки информационно-аналитических систем, предназначенных для извлечения знаний из растущего объема накапливаемых данных.

Изучение проблем и решение задач в этой области активно проводится в направлениях Business Intelligence (Интеллектуальный анализ данных) и Knowledge Management (Управление знаниями). В них 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 в этих направлениях положены в основу многих информационно-аналитических систем. Актуальным для их создания и последующего использования является разработка специализированных методик и программных инструментов, предназначенных для решения задачи кластеризации данных.

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.

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 и т.д.

На сегодняшний день в области кластерного анализа актуально решение следующих проблем: обоснованный выбор наиболее подходящего метода исследования; сложность оценки получаемых разбиений; отсутствие рекомендаций по применению методов в различных предметных областях; определение количества кластеров.

Прикладной областью диссертационной работы выбрана сфера брокерского обслуживания клиентов, для которой в настоящее время отсутствует достаточное количество практических рекомендаций по использованию существующих методов кластеризации, 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 на основе интеграции методов минимального остовного дерева и нечетких К-средних.

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, алгоритмов, нечетких множеств, графов, реляционных баз данных.

Научная новизна

Научную новизну работы составляют:

— методика адаптивной кластеризации фактографических данных;

— метод адаптивной кластеризации фактографических данных смешанного типа на основе интеграции методов минимального остовного дерева и нечетких К-средних (ADAKL), 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

Научное и народнохозяйственное значение работы состоит в разработке методики выполнения кластерного анализа фактографических данных и рекомендациях по использованию существующих и созданного методов кластерного анализа. Практическая ценность разработанного метода состоит в том, что он сокращает время проведения исследования. Предложенный в работе метод докластеризации позволяет проводить дополнительные исследования новых объектов без проведения общего анализа всех объектов, что приводит к сокращению временных затрат. Кроме этого, практическое применение результатов работы для исследуемой предметной области – брокерского обслуживания клиентов кредитной организацией, позволило решить задачу выделения существующих групп клиентов, находящихся на обслуживании.

Апробация работы

Основные положения диссертационной работы докладывались и обсуждались в 2006-2010 гг. на заседаниях комиссий по аттестации аспирантов и научных семинарах аспирантов и студентов МГТУ им. Н.Э. Баумана. Апробация работы проводилась на всероссийских и международных конференциях «Телематика 2009», «ИТ в образовании, науке и производстве 2009», «Бизнес-аналитика. Вопросы теории и практики. Использование аналитической платформы Deductor в деятельности учебных заведений 2010»; в рамках научной школы «Компьютерная графика и математическое моделирование»; на семинарах научно-образовательного кластера 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 составляет 185 страниц и содержит 30 таблиц, 21 схему, 137 источников, из них 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 методов:

CURE, BIRCH, MST, k-средних, PAM, CLOPE, самоорганизующиеся карты Кохонена (SOM), HCM, Fuzzy C-Means.

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).

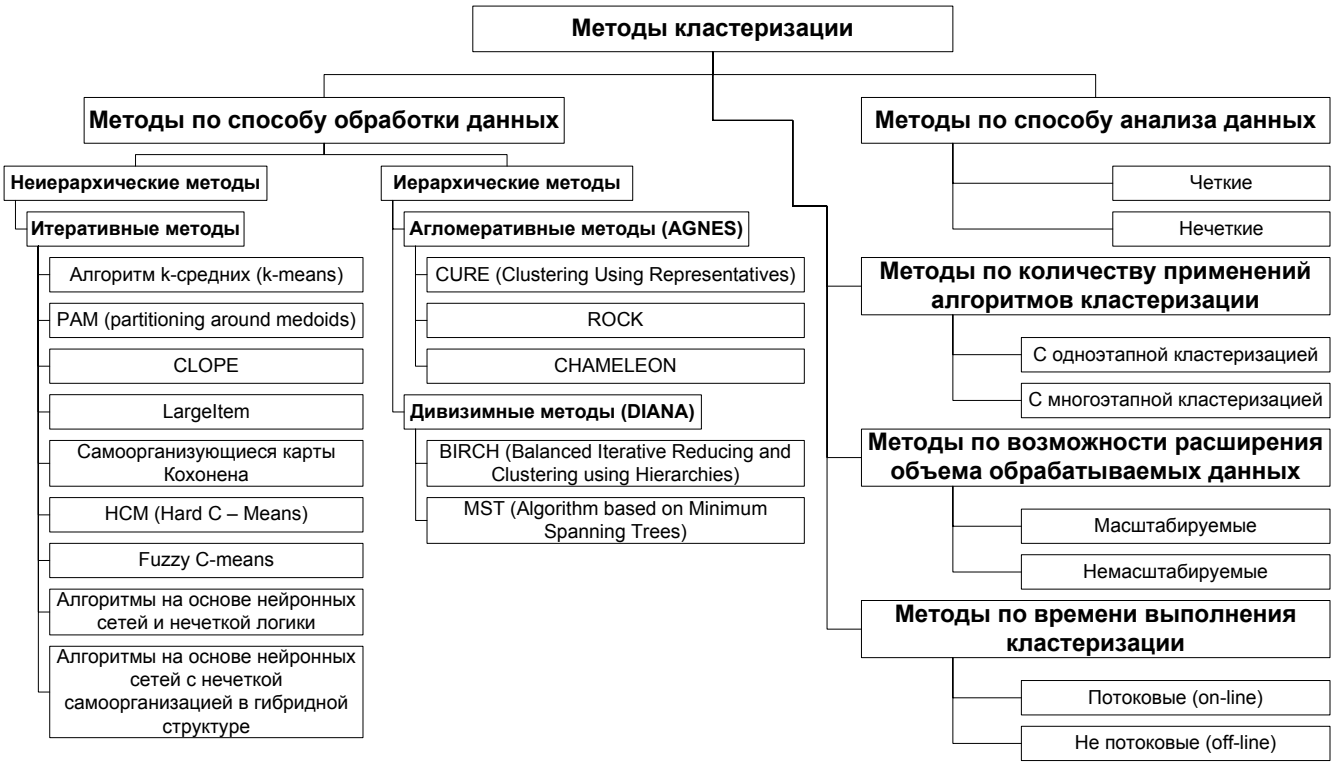

Рис. 1. Классификация методов кластеризации.

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 к 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.

На основе проведенного исследования предметной области выделен и формализован класс задач: количество исходных объектов: $KL = \left[KL_1, KL_2, ..., KL_i \right]$, $i \in \left[500; 50000 \right]$; количество значимых характеристик объектов: $KL_i = \left[k_1, k_2, ..., k_j \right], j \in \left[70; 150 \right]$; типы характеристик T: $T \in \left[\textit{числовые}, \textit{лингвистические} \right]$; форма получаемых кластеров – сложная, с пересечениями; количество кластеров N – результат исследования: $N \in \left[5; 30 \right]$.

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

Под методами адаптивной кластеризации в работе понимаются методы, входной параметр «Количество кластеров» которых определяется в результате предварительного исследования, включающего, например, оптимизацию локальных критериев оценки качества разбиения, стабилизацию получаемых центров кластеров и др.

В связи с тем, что на данный момент количество методов кластеризации велико, а существующих практических рекомендаций по их использованию недостаточно, была разработана методика адаптивной кластеризации фактографических данных (рис. 2), которая направлена на решение этой задачи.

Выборка исходных данных (этап 1, рис. 2) может производиться с помощью различных средств: путем построения регулярных запросов, ведения сведений в различных системах оперативного, аналитического учета и т.п. Полученная выборка подлежит исследованию с целью выявления значимых объектов/характеристик объектов (этап 2, рис. 2), 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 можно разработать контрольный пример (этап 3, рис. 2), который в дальнейшем будет использован для проверки действенности метода. Данный процесс необходимо выполнять с привлечением носителей экспертных знаний в исследуемой области.

На следующем этапе (этап 4, рис. 2) 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 (этап 5, рис. 2) и получение результата в виде конечного разбиения множества исходных объектов на кластеры.

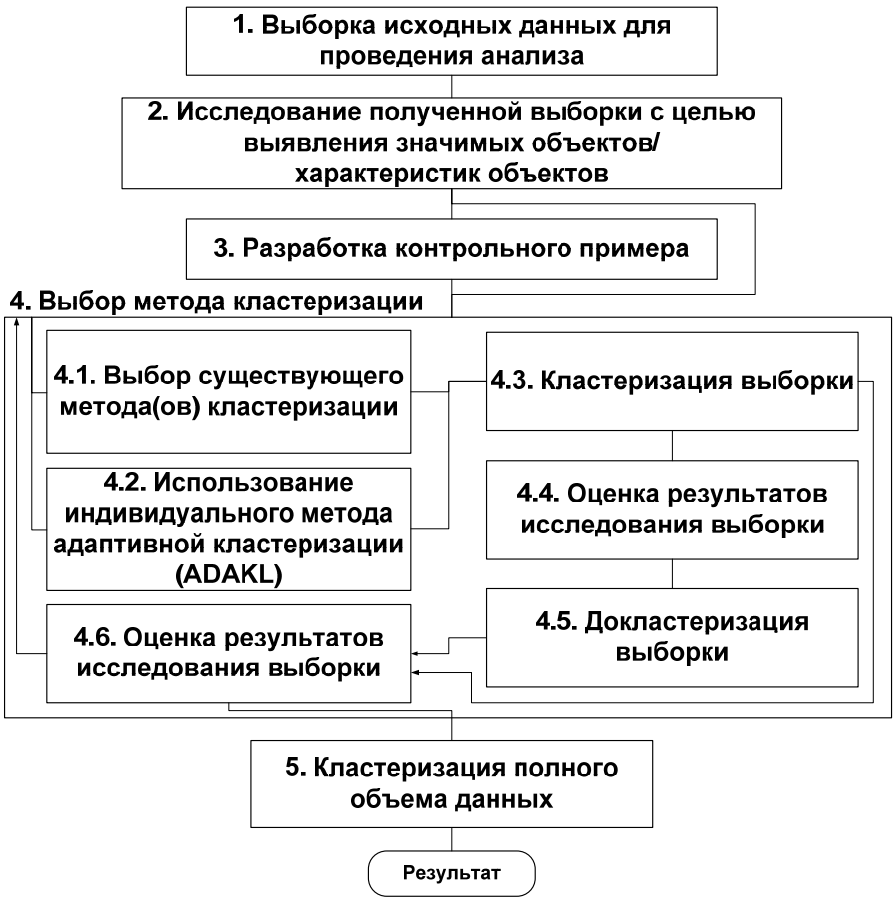

Рис. 2. Методика адаптивной кластеризации.

Выбор существующего метода кластеризации выполняется за три шага: 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 путем перебора существующих методов.

Выделено восемь критериев выбора метода: объем информации по отношению к времени обработки (Cr_1), размерность информации (Cr_2), типы атрибутов сущностей (Cr_3), чувствительность к равномерности информации (Cr_4), а также априорные (экспертные) представления о форме получаемых кластеров (Cr_5), их количестве (Cr_6) и перекрываемости (Cr_8). Для исследуемой предметной области установлены следующие приоритеты и значения выделенных критериев: 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*».

В результате анализа сформулированы рекомендации по использованию методов для исследования предметных областей и практических задач.

Сложностью кластеризации является необходимость получения такой 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 и др.

Аналитическая оценка сложности метода MST зависит от используемого алгоритма построения минимального остовного дерева: по алгоритму Борувки и Прима - $O\left[E \cdot Log\left(V\right)\right]$, по алгоритму Крускала - $O\left[E \cdot Log\left(E\right)\right]$, 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. Развивая идеи Загоруйко Н.Г., Елкиной В.Н., Айвазяна С.А., Бежаевой З.И., за основу метода в части первичного разделения объектов на кластеры взят метод MST, а для уточнения полученных кластеров используется 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 нечетких множеств - разделение объектов с лингвистическими атрибутами.

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 за счет использования предварительно настроенной словарной системы и теории нечетких множеств при определении соотношений между понятиями. Двухэтапность выполнения кластеризации и использование модифицированного критерия оптимальности позволяет повысить качество проводимой кластеризации.

Среди входных параметров метода присутствуют:

$U = \{u_1, u_2, ..., u_m\}$, где u_i – объекты кластеризации, m – количество объектов кластеризации, $i = \overline{1,m}$; $u_i = \{(Value_{i1}, t_1), (Value_{i2}, t_2), ..., (Value_{in}, t_n)\}$, где $Value_{ij}$ – значение j атрибута i объекта кластеризации, t_j – тип атрибута объекта кластеризации, n – количество атрибутов объекта кластеризации, $j = \overline{1,n}$; $K = \{K_1, K_2, ..., K_n\}$, где K_i – весовой коэффициент влияния атрибута объекта, $K_i \in [0;1]$; p – размазанность кластеров, $p \in (0;10]$; w – степень удаленности элементов, $w \in (0;1]$; q – максимальное количество кластеров, $q \le m$; k – текущее количество кластеров, $k \le q$.

Для устранения чувствительности к выбросам на первом этапе предлагается использование предобработки исходных данных для числовых атрибутов в виде линейной и статистической нормализации.

При вычислении информационных расстояний используются классические формулы вычисления (Евклидово расстояние, квадрат Евклидова расстояния, расстояние Чебышева), которые дополнены весовым коэффициентом K:

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

Для построения минимального остовного дерева на третьем этапе рекомендуется применять алгоритм Прима.

На основе построенной оптимизированной структуры объектов в виде дерева строится матрица нечеткого разбиения, которая обладает следующими характеристиками: $F = [\mu_{ij}], \mu_{ij} \in [0,1], i \le k, j = \overline{1,m}$, где μ_{ij} – степень принадлежности i объекта к j 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 V_i^k – центр кластера i для k итерации расчета, Avg – оператор вычисления среднего значения показателей объектов, входящих в кластер i, C_i^k – i кластер для k итерации расчета, $i = \overline{1,k}$, $j = \overline{1,m}$.

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 $Dist^k$ – матрица расстояний от объектов до центров кластеров для k итерации расчета, $i = \overline{1,k}$, $j = \overline{1,m}$, $Metric$ – способ определения информационного расстояния между объектами.

Рис. 3. Основные этапы метода адаптивной кластеризации ADAKL.

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

$Dist_{ij}^{k'} = \left\{Dist_{ij}^k \middle/ Max\left(Dist_{ij}^k\right) \middle| Max\left(Dist_{ij}^k\right) \neq 0; Dist_{ij}^k \middle| Max\left(Dist_{ij}^k\right) = 0\right\}$, где $Dist_{ij}^{k'}$ – нормализованная матрица взаимных расстояний от объектов до центров кластеров для k итерации расчета, $i = \overline{1,k}$, $j = \overline{1,m}$.

При соотнесении объектов к кластерам в соответствии со степенью удаленности элементов кластера 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:

$$\mu_{ij} = \left\{ \mu_{ij} \Big/ \sum_{i=1}^{k} \mu_{ij} \left| \sum_{i=1}^{k} \mu_{ij} \neq 0; \mu_{ij} \right| \sum_{i=1}^{k} \mu_{ij} = 0 \right\}, \text{ где } i = \overline{1,k},\ j = \overline{1,m}.$$

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

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

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

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

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:

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: $O_{Onm} = \underset{i=1,q}{MAX}\left(O^i\right)$, где $i = \overline{1,q}$.

Метод ADAKL 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 (рис. 4), которая определяет необходимость повторного запуска исследования полного массива данных и, в случае отсутствия признаков появления новых значимых групп объектов, осуществляет распределение новых объектов по имеющимся кластерам. Процесс докластеризации интегрируется с основным методом, однако может выполнять некоторые этапы независимо от основного алгоритма.

Необходимость в докластеризации подтверждается результатами эмпирических исследований, по результатам которых выявлено, что наиболее трудоемким этапом метода является построение минимального остовного дерева. 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.

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

$$O_1 = \sqrt{\frac{\sum_{i=1}^{r} \left\| A_i - Avg\left[A\right] \right\|^2}{r}},\ O_2 = \sqrt{\frac{\sum_{i=1}^{o} \left\| B_i - Avg\left[B\right] \right\|^2}{o}},\ \left| O_1 - O_2 \right| \le \delta,$$ 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; δ – доверительный интервал; $\left\| A_i - Avg\left[A\right] \right\|, \left\| B_i - Avg\left[B\right] \right\|$ – 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 K основного алгоритма.

Пороговое значение, обозначающее подобность обоих наборов данных, является входным параметром метода докластеризации и инициализируется в методе ADAKL.

Распределение элементов расширяющего множества (B) по вычисленному расстоянию до ближайших k объектов (в соответствии с критерием $O_{Onm} = \underset{i=1,q}{MAX}\left(O^i\right)$) из расширяемого множества (A) выполняется следующим образом:

$\mu_{ij} = \left(1 - Dist_j^{Norm}\right) * \mu_{il} = \left[1 - Dist_{lj} \Big/ Max\left(Dist_{sj}\right)\right] * \mu_{il}$, где $i = \overline{1,k}$ – номер кластера множества A; $j = \overline{1,o}$ – порядковый номер элемента из множества B; $s \in \left[1, r\right]$ – порядковый элемента из множества A из ближайших k объектов; l – порядковый номер бли-

жайшего элемента из множества A для элемента из множества B, $l \in [1,r]$; $Dist_{lj} = \|B_j - A_l\|$ – расстояние между ближайшими элементами из множеств A и B; $Max(Dist_{sj})$ – максимальное расстояние от элемента из множества B до элемента из множества A; μ_{il} – степень принадлежности элемента l из множества A к кластеру i.

Рис. 4. Докластеризация исходных данных.

В результате расчета 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 границу степени принадлежности объектов кла-

стерам; 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. 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 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.

Во второй половине главы описываются три основных и одна дополнительная серии по пятьдесят экспериментальных исследований для оценки работоспособности ADAKL в сравнении с другими алгоритмами (SOM, k-средних, ADAKL). Исследуемые массивы данных имеют следующие характеристики:

Табл. 1. Сводная таблица исследуемых данных.

Характеристика / **Исследование**	**Общее кол-во исх. данных (шт. записей)**	**Кол-во атрибутов**		
		Числовые (шт.)	**Текстовые (шт.)**	**Общее (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 результатов кластерного анализа выполнена на основе показателей выполненной кластеризации с помощью индекса истинности разбиения:

$O = \frac{r}{n} * \begin{cases} q/k, q \le k \\ k/q, q > k \end{cases}$, где q – количество кластеров по итогам кластеризации; r – количество элементов, правильно распределенных по соответствующим кластерам; k – исходное количество кластеров; n – количество объектов кластеризации.

В соответствии с полученной итоговой оценкой (табл. 2) 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 раз-

работанного метода достигается использованием математического аппарата нечетких множеств и внутренних словарей системы при определении информационных расстояний между объектами.

Табл. 2 Средневзвешенная оценка разбиений.

Оценка / Метод	Средневзвешенная оценка			
	Без задания кол-ва кластеров	**С заданным количеством кластеров (без учета лингвистических атрибутов)**	**С заданным количеством кластеров (с учетом лингвистических атрибутов)**	**Итоговая**
1	0.7913	0.9150	0.9237	0.8767
2	-	0.8232	-	0.8232
3	0.9762	0.9981	0.9990	0.9911

Для проверки аналитической оценки метода проведено пять серий по пятьдесят нагрузочных экспериментов на основе эмпирических данных, которые подтвердили аналитическую оценку метода.

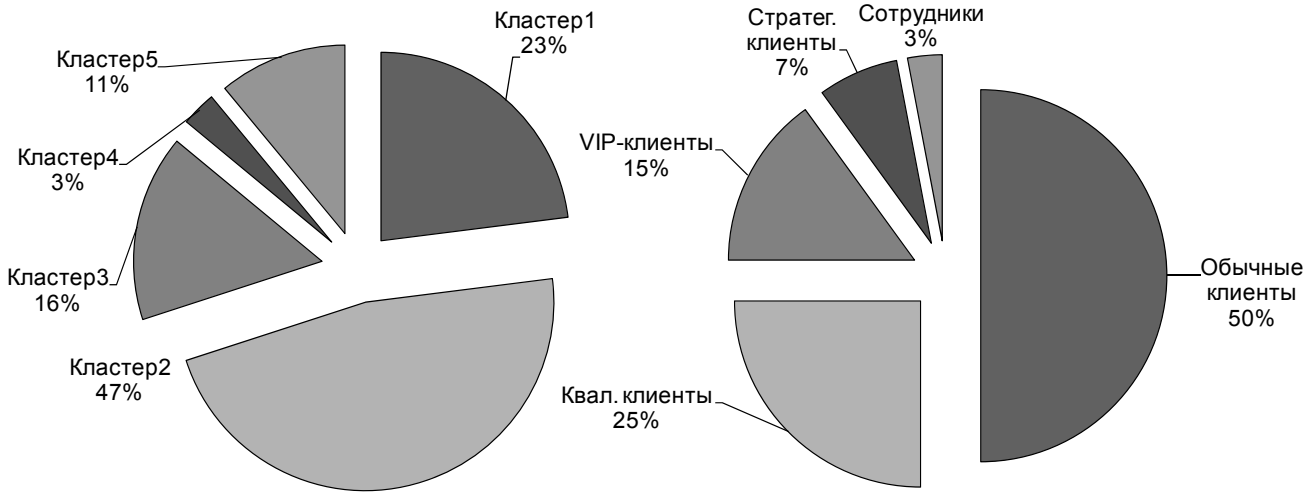

Рис. 5. Распределение клиентов по группам.

С помощью ПР решена практическая задача в области брокерского обслуживания по выделению групп клиентов и определена их доля от общего количества клиентов (рис. 5). 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. Это 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 выполнения кластерного анализа данных в выбранной предметной области, а также повышают эффективность и качество кластеризации за счет интеграции методов.

Основные результаты диссертационной работы

1. Проведено исследование существующих методов и подходов ИАД, используемых для кластеризации фактографических данных.

2. 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, кластеризация полного объема данных.

3. На основе литературных источников для выбора метода кластеризации выделено восемь критериев и разработаны рекомендации.

4. Разработан метод адаптивной кластеризации (ADAKL) на основе интеграции методов минимального остовного дерева и нечетких К-средних, определяющий количество кластеров с помощью локального критерия, обладающий двухэтапностью, нечеткостью при распределении объектов по кластерам, возможностью использования объектов с разными типами атрибутов, приемлемым временем работы и конечностью результата.

5. 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.

6. 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.

7. Разработанный метод ADAKL реализован в виде программного решения и проведены серии экспериментов, которые подтверждают аналитическую оценку и состоятельность в сравнении с имеющимися методами (k-средних, SOM).

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+», 2005. – 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.: НОК «CLAIM», 2006. – Выпуск 8. – С. 130-142.

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.

7. Нейский, И.М., Филиппович, А.Ю. Сегментация клиентов брокерского обслуживания / И.М. Нейский, А.Ю. Филиппович // Бизнес-аналитика. Вопросы теории и практики. Использование аналитической платформы Deductor в деятельности учебных заведений: сборник материалов межвуз. науч. практ. конф. – Рязань: Лаборатория баз данных, 2010. – С. 102-111.

8. Нейский, И.М. Экспериментальные исследования адаптивной кластеризации фактографических данных / И.М. Нейский // Материалы научной межвузовской конференции преподавателей, аспирантов, молодых ученых и специалистов «Печатные средства информации в современном обществе (к 80-летию МГУП)». Секция «Электронные средства информации в современном обществе. Сб. тез. докл. – М.: 2010. – С. 55-58.

Подписано в печать 22.10.2010 г.
Формат бумаги 60×90. 1/16. 1 п.л.
Тираж: 100 экз. Заказ № 354

Типография Aegis-Print
115230, Москва, Варшавское шоссе, д. 42
Тел.: 543-50-32
www.autoref.ae-print.ru