Сегментация клиентов брокерского обслуживания

***Нейский И.М.**, аспирант Московского государственного технического университета им. Н.Э. Баумана*
***Филиппович А.Ю.**, к.т.н., доцент Московского государственного технического университета им. Н.Э. Баумана*

Большинство современных предприятий используют в своей деятельности информационные системы, хранилища данных, в которых собираются данные по бизнес – процессам компании. Объемы накапливаемой информации увеличиваются с течением времени, поэтому актуальной задачей в развитии компании является переход от анализа тенденций текущих показателей деятельности предприятия к более комплексном подходу «извлечения знаний» из имеющихся данных в целях выявления закономерностей.

Изучением проблем и созданием решений в этой области активно занимаются направления Интеллектуального анализа данных (Business Intelligence) и Управления знаниями (Knowledge Management), в рамках которых выделяются поднаправления Выявление знаний в базах данных (Knowledge Discovery in Databases), Анализ фактографических данных (Data Mining), Анализ неструктурированных данных (Text Mining) 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: классификация, регрессия, кластеризация, выявление ассоциаций, выявление последовательностей, и 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.

Примером организаций, для которых актуальна проблема анализа накопленной информации, являются финансовые компании – профессиональные участники, работающие на фондовом рынке, которые привлекают клиентов на брокерское обслуживание. На сегодняшний день в России существует более 60 крупных компаний со среднемесячным оборотом около 800 миллионов долларов США [1]. Основным показателем эффективности работы в данном направлении является объем клиентских оборотов и комиссионных сборов за совершаемые от их имени и за их счет операции, поэтому для успешного развития брокерского обслуживания необходимо увеличивать количество клиентов и/или их обороты, на основе которых, как правило, определяется сумма комиссионного вознаграждения. Ввиду того, что каждый клиент по-своему уникален (желания, возможности, предпочтения, стратегия и т.п.), то для его привлечения и создания заинтересованности от компании требуется существенная гибкость. Так как отвечать интересам каждого клиента со стороны крупной компании, обслуживающей более 10 000 клиентов, практически невозможно при текущем штате сотрудников компании, которые сопровождают заключение и оперативную обработку этих операций, поэтому для обеспечения дальнейшего развития компании проводится анализ клиентской базы с целью выделения характерных групп клиентов. По результатам данного исследования для полученных групп клиентов разрабатываются индивидуальные тарифы, условия обслуживания и т.д.

Данный подход подтверждает свою эффективность даже в кризисный период, так как при падении на мировых финансовых рынках экономических показателей происходит отток капитала и клиентов. Создание индивидуальной продуктовой линейки и снижение необходимости значительного расширения штата сотрудников для сопровождения новых операций клиентов ведет к снижению издержек, а значит и к снижению тарифов при их обслуживании, что привлекает новых и расширяет количество операций существующих клиентов. Кризис – это не только спад показателей в различных отраслях экономики, но и возможность достичь более значимых результатов за счет повышения собственной эффективности, поэтому появляется необходимость использовать дополнительные, ранее не используемые ресурсы, которые сосредоточены в компании – внутреннее информационное поле (аналитики, эксперты, накопленная информация об операциях, клиентах и т.д.).

Главная особенность анализа этой области в том, что его необходимо проводить на регулярной основе, чтобы сохранить конкурентные преимущества на рынке данного вида услуг. Учитывая динамику роста клиентской базы, использование ранее применяемых методов с привлечением только человеческих ресурсов становится невозможно, так как объем информации для анализа также возрастает. На данный момент известно более 100 методов кластеризации, поэтому для перехода на использование машинных методов необходимо осуществить выбор метода или методов из существующих либо разработать собственный метод с учетом особенностей этой области. Анализ существующих решений и методов [2] показал, что на текущий момент не существует специализированных или успешно примененных универсальных методов для решения описанной задачи. Еще одной проблемой в данной области является оценка качества получаемого результата и выбор количества групп – кластеров, которое является входным параметром для большинства алгоритмов.

В связи с тем, что на данный момент не существует достаточного количества практических рекомендаций по применению существующих методов в данной предметной области и количество методов достаточно велико, была разработана методика адаптивной кластеризации, которая направлена на решение этой задачи. Данная методика, состоящая из четырех этапов, позволяет осуществить выбор метода кластерного анализа и получить конечное разбиение множества исходных объектов на кластеры. На основе методики получено, что для решения задачи разбиения клиентов брокерского обслуживания необходимо разработать новый метод адаптивной кластеризации, в котором количество кластеров является результатом исследования.

После проведенного анализа для решения поставленной задачи из инструментов выполнения кластеризации были выбраны: теория графов и нечеткая логика. Определяющими факторами в выбранной комбинации является способность при использовании графов выделять кластеры произвольной формы и оптимальной структуры, а при использовании математического аппарата нечеткой логики решается задача разделения объектов с лингвистическими атрибутами. За основу для нового метода в части первичного разделения объектов на кластеры взята идея метода MST [3], использующего минимальные остовные деревья, и идея метода Fuzzy C-means [4]. На базе этих методов разработан метод ADAKL (рис. 1), который является двухэтапным и использует оценочную функцию разбиения, повышающую качество проводимой кластеризации [5]. Вычисление глобального критерия делает алгоритм кластеризации во

много раз быстрее, чем при использовании локального критерия при парном сравнении объектов.

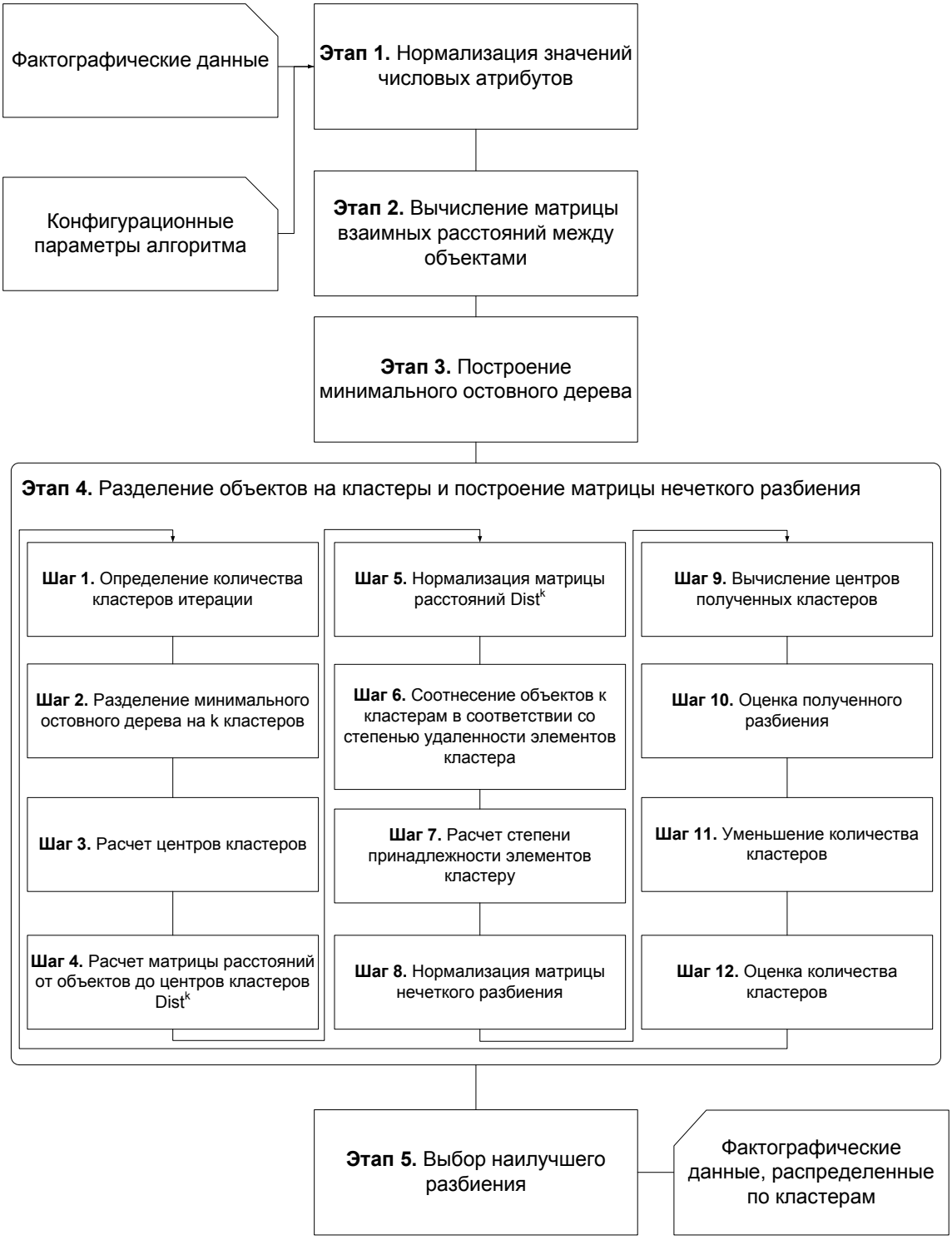

Рис. 1. Метод ADAKL.

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.

Для устранения чувствительности к выбросам на первом этапе предлагается использование предобработки исходных данных через линейную или статистическую нормализацию. При вычислении информационных расстояний между объектами используются классические метрики, доработанные для использования в методе: Евклидово расстояние, квадрат Евклидова расстояния, расстояние Чебышева. Для построения минимального остовного дерева используется алгоритм Прима, т.к. он имеет наименьшую аналитическую сложность по сравнению с алгоритмами Крускала и Борувки [6].

Оценка качества в методе ADAKL выполняется на основе локального критерия с использованием полученных центров кластеров:

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

k - количество кластеров;

m – количество объектов кластеризации;

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

$\mu^p{}_{ij}$ - степень принадлежности $i^{го}$ объекта к $j^{му}$ кластеру;

p – размазанность кластеров;

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

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

Данная оценка нацелена на выделение кластеров с наименьшими взаимными расстояниями и наибольшим количеством элементов в кластере по отношению к общему количеству кластеров, стремится к уменьшению количества итоговых кластеров и нацелена на минимизацию взаимных расстояний между полученным центром кластера и элементами с учетом степени принадлежности.

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.

Вместе с тем 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.

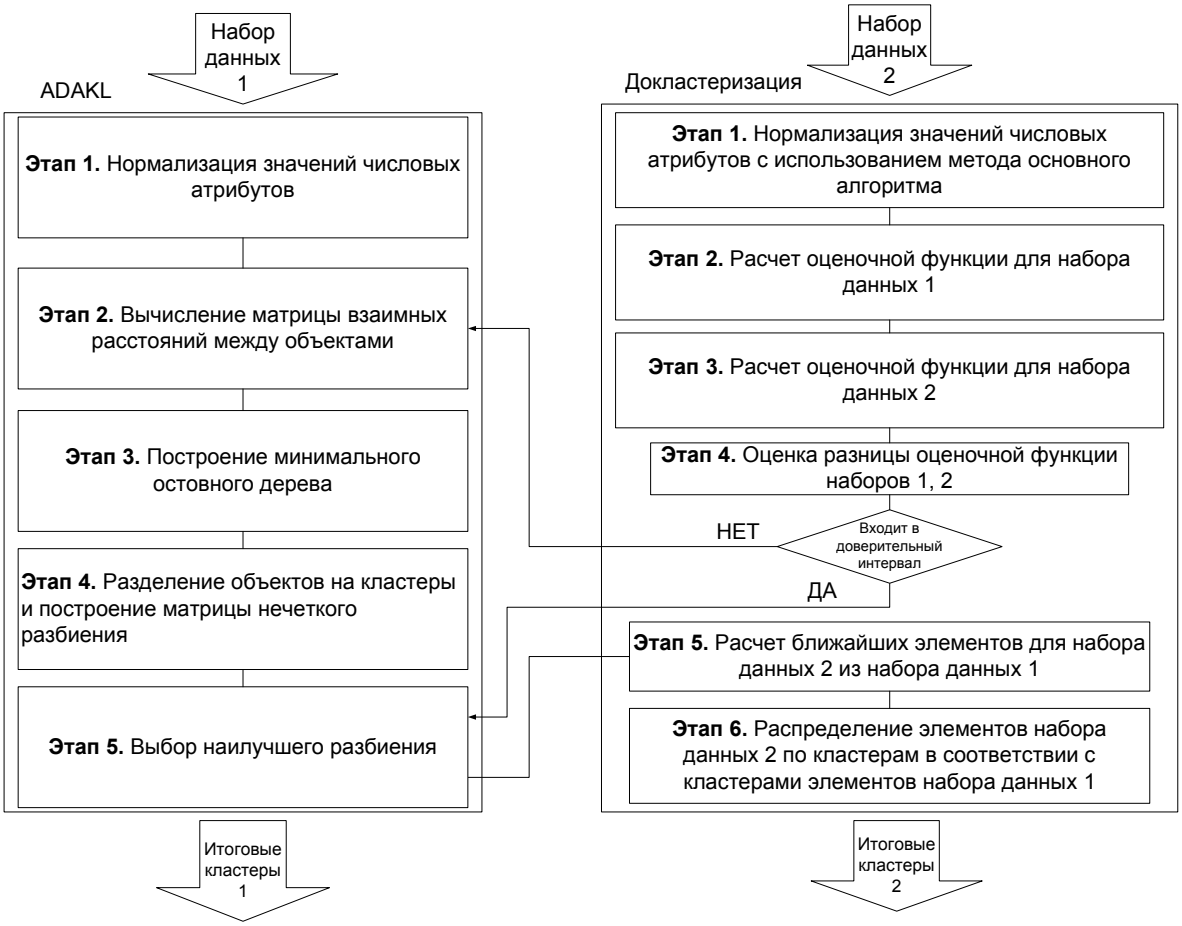

Рис. 2. Докластеризация дополнительного набора данных.

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. 2).

Необходимость в докластеризации подтверждается результатами эмпирических исследований, по результатам которых выявлено, что наиболее трудоемким этапом метода является построение минимального остовного дерева [5]. 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.

Для оценки работоспособности ADAKL в сравнении с другими алгоритмами были проведены три основных и одна дополнительная серии экспериментов:

- выделение секторов инвестирования на основе анализа показателей финансовых инструментов;
- выделение групп клиентов на основе статистических данных о деятельности клиентов за период;
- выявление категорий финансовых инструментов для оценки эффективности операций;
- выделение классов автомобилей на основе данных о максимальной скорости, цвете кузова, воздушном сопротивлении, массе.

Исследование производилось на трех методах (самоорганизующиеся карты Кохонена, алгоритм k–средних и разработанный метод ADAKL) с помощью аналитической платформы Deductor Academic и разработанного программного решения, в котором реализован метод ADAKL. По результатам исследования составлена сводная таблица с усредненными оценками разбиений (таблица 1).

Таблица 1

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**
Карты Кохонена	0.7913	0.9150	0.9237	0.8767
k–средних	-	0.8232	-	0.8232
ADAKL	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.

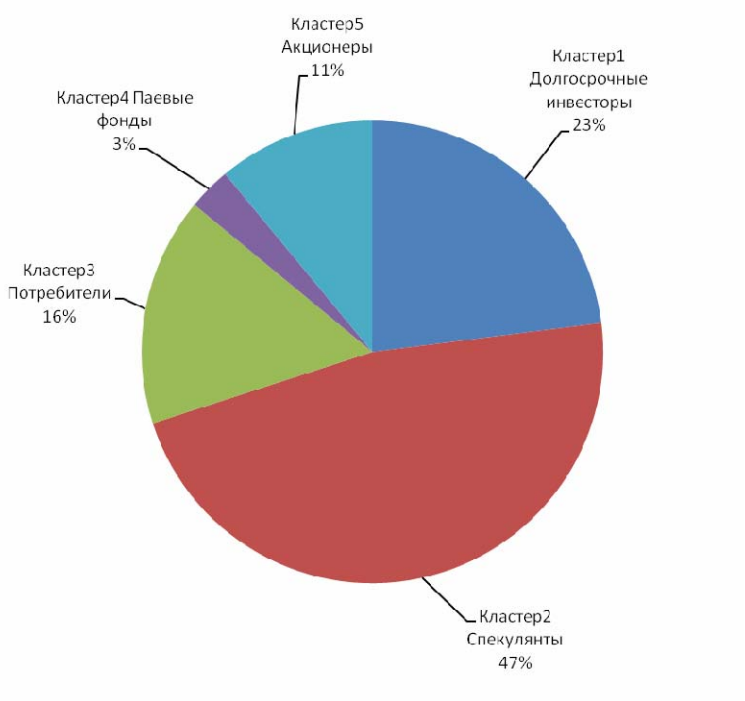

Рис. 3. Распределение клиентов по группам.

На основе метода ADAKL были выделены группы клиентов и определение их доли от общего количества клиентов (рис. 3). Последующий анализ экономических показателей полученных групп объектов позволил дать названия кластерам, 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. Дополнительная информацию о методике адаптивной кластеризации представлена в публикации [5] и на сайте научно-образовательного кластера CLAIM (http://philippovich.ru).

Список литературы

1. Прытин Д. Крупнейшие брокеры России // Источник: http://rating.rbc.ru.

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

3. Speer N., Merz P., Spieth C., Zell A. Clustering Gene Expression Data with Memetic Algorithms based on Minimum Spanning Trees. // University of Tubingen, Center for Bioinformatics. Источник: fs.informatik.uni-tuebingen.de.

4. Штовба С. Д. Введение в теорию нечетких множеств и нечеткую логику. // Источник: matlab.exponenta.ru.

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

6. Рыбаков Г. Построение минимального остовного дерева (алгоритмы Крускала, Прима, Борувки). - 2005. Источник: http://rain.ifmo.ru.