ISSN 1992-6502 (Print) Вестник УГАТУ ISSN 2225-2789 (Online)
2021. Т. 25, № 3 (93). С. 132–144 http://journal.ugatu.ac.ru

УДК 004.65

ПРОГРАММНОЕ РЕШЕНИЕ ЗАДАЧИ СЕГМЕНТИРОВАНИЯ НА ОСНОВЕ ИНТЕЛЛЕКТУАЛЬНЫХ ТЕХНОЛОГИЙ

А. В. Якупова[1], О. Н. Сметанина[2], Е. Ю. Сазонова[3]

[1] ajsylu.yakupowa@yandex.ru, [2] smoljushka@mail.ru, [3] ekaterina_rassadnikova@mail.ru

[1]Региональный центр IBS в Уфе
[2,3] ФГБОУ ВО «Уфимский государственный авиационный технический университет» (УГАТУ)

Поступила в редакцию 07.06.2021

Аннотация. Рассматриваются вопросы решения задачи сегментации объектов с учетом их значимых характеристик, что, с одной стороны, позволяет сузить поиск решений, с другой стороны, при анализе данных и интерпретации результатов анализа, включая сегментацию, формализовать рекомендации. Приводится постановка задачи сегментации, структура решения, а также осуществляется выбор методов для проведения сегментации и разработка алгоритмов. Особое внимание уделено проблеме сбора данных и их структурирования для решения задачи, обоснованию использования для сбора гибридного метода и выбору технологий для реализации сбора данных. Авторами предлагается на основе формализованных рекомендаций разработать программное решение.

Ключевые слова: задача сегментирования; интеллектуальные технологии; методы машинного обучения; структурирование данных; гибридный метод сбора данных.

ВВЕДЕНИЕ

Сегментация рассматривается в настоящее время как этап решения целого круга задач. В частности, она используется при обработке изображений с целью упрощения его представления. Другим примером может служить сегментация в маркетинге, позволяющая разделить объекты на сегменты и строить поддержку решений с учетом знаний о сегментах.

Во втором случае принципиальным видится выбор факторов, по которым может быть проведена сегментация. Так, на рынке недвижимости факторы могут быть связаны с целями потребителей, а именно, для улучшения жилищных условий или поиска недвижимости как инвестиционного проекта. По отношению к рынку как к системе факторы могут быть внешними и внутренними, сами на рынке могут относиться к «первичному» или «вторичному» рынку.

Рассмотрением вопросов сегментации рынка занимаются многие специалисты, как в России, так и за рубежом. Среди них: Е. И. Тарасевич, Г. М. Стерник, И. А. Бузова, Н. В. Васильева, С. Н. Максимова, В. А. Горемыкин, Д. Л. Волкова, Дж. Найта (J. Knight), Д. Стоукена (D. Stoken), В. Л. Борна (W. L. Born), С. А. Пирр (S. A. Pyhrr), Дж. Р. Вебб (J. R. Webb) [1–4].

Авторы статьи в ряде работ также затрагивали вопросы сегментации при организации информационной поддержки управленческих решений [5–7].

Кроме того, для той или иной задачи сегментации разработаны специальные программные решения: анализ и сегментация товара по степени важности (программа для ЭВМ

Работа поддержана грантом РФФИ 19-07-00709 и выполнена в рамках государственного задания № FEUE-2020-0007.

INDEEPA); выделение сегментов (кластеров) среди регионов по выручке от рекламной деятельности; автоматическое построение диаграммы выделенных сегментов среди регионов (программа для ЭВМ «сегментация рынка рекламы») и др. Непосредственно для анализа и совершения сделок в области недвижимости существует множество программных решений, которые отличаются функционалом (одна группа позволяет анализировать объекты недвижимости, другая – специализируется только на поиске недвижимости и совершении сделок на рынке), например, ДомКлик и Циан.

Несмотря на то, что вопросами сегментации для различных рынков занимается широкий круг специалистов, как теоретиков, так и практиков, а также имеется ряд программных решений в этой области, специфика постановки задачи, выбор факторов для проведения сегментации, специфика факторов, источников данных и вопросы автоматизации их сбора требуют разработки или адаптации моделей и методов решения задач. И, как следствие, разработки программного решения на их основе.

В статье ставится задача сегментации на примере рынка недвижимости, рассмотрены вопросы выбора факторов автоматического сбора структурированных данных из распределенных источников (например, Авито, ЦИАН), а также моделей и методов для ее решения, формализации требований к разрабатываемому программному решению и его реализация.

ПОСТАНОВКА ЗАДАЧИ СЕГМЕНТИРОВАНИЯ ОБЪЕКТОВ (НА ПРИМЕРЕ РЫНКА НЕДВИЖИМОСТИ) И ЕЕ ПРОГРАММНОЙ РЕАЛИЗАЦИИ

Задача сегментации является одним из базовых этапов при проведении анализа и решается для объектов рынка недвижимости с учетом факторов, влияющих на ценообразование объектов.

Формальная постановка для программной реализации задачи анализа рынка недвижимости, включая его сегментацию, осуществляется с учетом требований, предъявляемых пользователем, и на основе собранных данных об объектах недвижимости с URL-страниц с целью определения сегментов объектов недвижимости, графического отображения сегментов и интерпретации результатов сегментации с описанием характеристики исследуемого кластера (рис. 1).

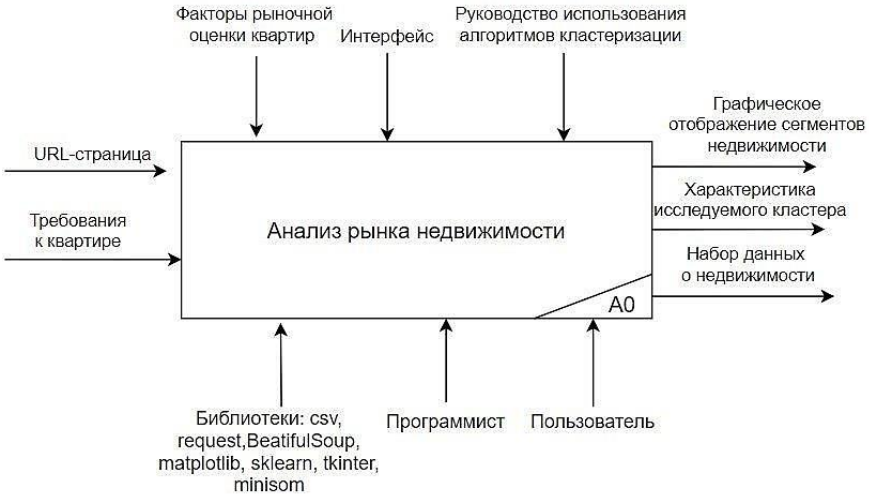


Рис. 1. Формальная постановка задачи

Математическая постановка задачи сегментации. Дано: конечное множество объектов – $P = \{p1, p2, \ldots, pn\}$, каждый из которых описывается множеством характеристик $F_{pi} = \{f1_{pi}, f2_{pi}, \ldots, fm_{pi}\}$, а также определяющий правило сегментации предикат pr.

Найти: разбиение P на $S = \{s1, s2, \ldots, sk\}$ сегментов, когда предикат pr является истиной, т.е. Seg: $pi \xrightarrow{pr} sj$, где $sj \in Ssk \in S$ и $sj \bigcap sk = \emptyset$.

При описании объекта недвижимости следует обратить внимание на множество характеристик объекта недвижимости (рис. 2).

Рис. 2. Характеристики объекта недвижимости

Для анализа рынка недвижимости необходимо получить реальные данные об объектах недвижимости в проверенных источниках, сохранить результат сбора информации в удобном виде для чтения, при этом разделить выборку на 4 подгруппы (первичный рынок, вторичный рынок, аренда посуточно, долгосрочная аренда), провести предобработку каждой выборки, выделить основные значащие характеристики, разбить полученную информацию наилучшим из рассматриваемых алгоритмов кластеризации и вывести результаты исследования в удобном для пользователя формате.

Необходимую информацию об объектах недвижимости (первичного и вторичного рынков) можно найти на сайте «Авито»: цена, адрес, тип, этаж, количество этажей в доме, район, площадь квартиры, площадь кухни, жилая площадь и тип отделки, информация об инфраструктуре дома, варианты оплаты.

Для автоматизированного сбора информации об объектах недвижимости, точнее для извлечения данных из веб-страницы в структурированном виде, предложено написать скрипт. Для обработки и анализа собранных данных, включая задачи сегментирования рынка недвижимости, представления характеристик выделенных сегментов с оценкой средней стоимости объектов недвижимости разработать программное решение, включающее математическое и информационное обеспечение перечисленных задач.

Программное решение должно обладать следующим функционалом: ввод характеристик объекта недвижимости (квартиры, включая и характеристики дома); возможность удаления введенных данных; очищение прошлых результатов программы; разделение объектов недвижимости на сегменты путем применения методов кластеризации; отображение результатов анализа недвижимости (графическое отображение сегментов недвижимости и характеристика кластера введенной квартиры); осуществление поиска соответствующих объектов недвижимости согласно предъявляемым требованиям; отображение найденных объектов недвижимости в виде таблицы с данными.

ПОДХОД К РЕШЕНИЮ ЗАДАЧИ СЕГМЕНТИРОВАНИЯ ОБЪЕКТОВ НЕДВИЖИМОСТИ

Решение задачи основано на последовательном выполнении операций сбора информации об объектах недвижимости и ее структурирования, а также непосредственного анализа данных, включающего сегментирование (рис. 3).

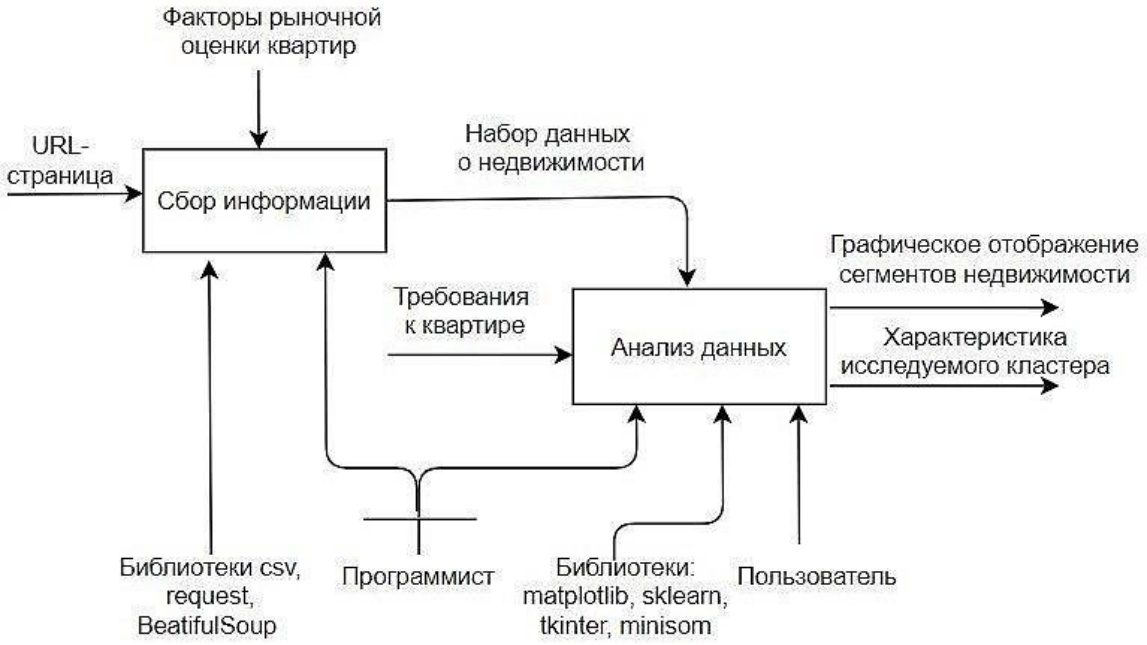


Рис. 3. Декомпозиция 1 уровня

Информация из сети Интернет может быть использована для широкого класса задач, к таким задачам можно отнести задачи интеллектуальной обработки данных, например, задача классификации новостей с целью выявления неправдивых (фейковых новостей), задача сегментации недвижимости, задачи оптимизации (например, задача маршрутизации транспорта с использованием геоинформационных картографических сервисов и др.). Поскольку разработчики сайтов могут как соблюдать правила HTML-разметки, то есть выделять блоки, придерживаться одной структуры разметки, так и не соблюдать данные правила или менять структуру, то, как правило, приходится иметь дело с полуструктурированной информацией. Часто информация в сети Интернет может быть и неструктурированной. В случае, когда структура сайта не единообразна, информацию извлечь становится сложнее. Актуальность и сложность сбора информации в сети Интернет привели к необходимости решения задачи разработки технологий, позволяющих извлекать необходимые данные.

Данные из Интернет-ресурсов можно получить следующими методами: ручной, гибридный или автоматический. В ручном методе основную работу выполняет человек, то есть он выполняет всю цепочку действий самостоятельно: находит необходимую информацию, копирует и сохраняет в удобном для обработки виде. Данный метод имеет ряд недостатков: он пригоден только для небольших объемов данных, качество собранной информации зависит от ответственности и компетентности человека.

Гибридный метод – метод, в котором человек несколько облегчает работу извлечения информации, используя вспомогательные средства, такие как различные библиотеки, Headless-браузеры, SaaS решения (рис. 4).

Под «автоматическим» методом понимается получение и структурирование данных, которое производится с помощью систем автоматического распознавания данных и их структур.

Наиболее близкими к данному типу являются технологии поисковых гигантов и больших аналитических агентств.

Для сбора данных о недвижимости был выбран гибридный метод, а именно использовалась библиотека lxml/html для языка Python.

Анализ данных включает в себя этапы предобработки, сегментированная (кластеризации) и графического отображения кластеров. Для сегментирования вторичного жилья предложено использовать самоорганизующиеся карты Кохонена, для первичного жилья – k-means.

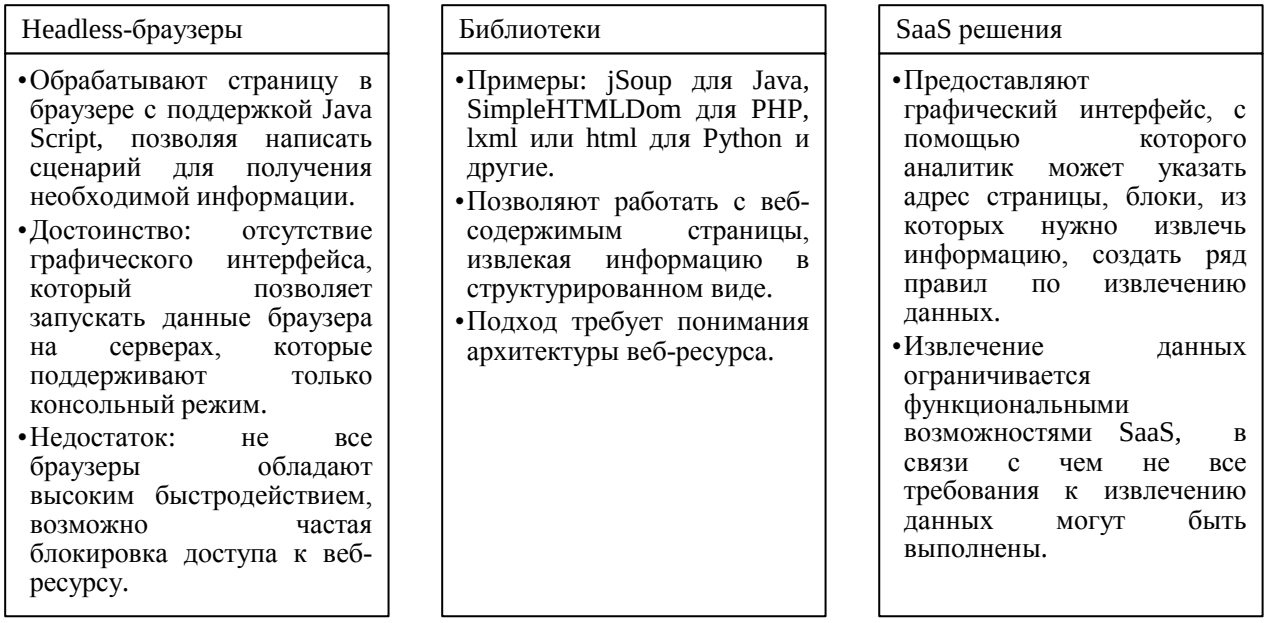


Рис. 4. Технологии гибридного метода

МЕТОДЫ И АЛГОРИТМЫ ПРЕДЛАГАЕМОГО ПОДХОДА

Алгоритм парсинга. Под термином «парсинг» понимается автоматизированный сбор неструктурированной информации, ее преобразование и выдача в структурированном виде. Парсинг позволяет ускорить процесс сбора данных и избежать ошибок из-за человеческого фактора. Как правило, парсеры пишут отдельно для каждого сайта с учетом его структурных и технических особенностей. При сборе данных с помощью парсинга могут возникнуть следующие затруднения и ограничения: по user-agent (запрос, в котором программа сообщает о себе сайту); по robots.txt (программа, блокирующая поисковых роботов); по IP-адресу (если в течение длительного времени с одного адреса поступает много запросов, сайт блокирует данные); по капче (если действия похожи на автоматические, то выводится капча) и др. Достоинствами парсинга являются: высокая скорость сбора данных в любом режиме работы; возможность задавать различные параметры; минимум ошибок по невнимательности человека; выполнение регулярной проверки по заданному времени; представление данных в любом формате; равномерное распределение нагрузки на сайт.

Пошаговое описание алгоритма парсинга для сбора информации об объектах недвижимости включает 6 шагов.

Шаг 1. Построение запроса для получения информации.

Шаг 2. Осуществление запроса и получение ответа на запрос.

Шаг 3. Обработка ответа.

Шаг 4. Получение ссылки каждой квартиры с одной страницы.

Шаг 5. Осуществление перехода по каждой полученной ссылке с целью извлечения данных со страницы: заголовок квартиры, цена, адрес, тип дома, этаж, количество этажей в доме, площадь квартиры, площадь кухни, жилая площадь, тип отделки.

Шаг 6. Структурирование необходимой информации.

После того как информация со страниц будет собрана в массив данных, необходимо провести предобработку данных.

Методы предобработки данных. Качество анализа данных очень сильно зависит от предобработки данных. Собранные данные должны отражать генеральную совокупность и быть пригодными для дальнейшей обработки.

Для получения качественных данных должны быть выполнены следующие этапы предобработки данных:

1. Оценка репрезентативности данных. Для этого необходимо проверить пропорции категориальных данных в выборке, для числовых данных должны быть проанализированы основные статистики выборки, такие как математическое ожидание, среднеквадратическое отношение, коэффициенты асимметрии и эксцесса, и др.

2. Оценка центрального положения, то есть оценка типичных значений для числовых признаков, для категориальных данных определить часто и мало встречаемые значения.

3. Очистка данных. На данном этапе необходимо избавиться от ошибочных значений, обработать отсутствующие записи, удалить дубликаты, проверить формат данных, оценить отсутствующие данные и принять решения по использованию таких данных.

4. Преобразование категориальных данных в числовой формат без потери смысла. Для этого необходимо категориальные значения преобразовать в отдельные признаки – фиктивные переменные. При использовании данного способа увеличивается размерность выборки, в таком случае следует преобразовать данные в разряженную матрицу и использовать специальные методы для обработки данных.

5. Приведение данных к одному формату. Так как в наборе данных могут находиться признаки, сильно различающиеся по значениям, и это будет приводить к тому, что одни признаки будут перевешивать другие, необходимо преобразовать данные к одному диапазону, чтобы каждый признак вносил одинаковый вклад, то есть не зависел от масштаба и не оценивался по важности значения. Недостатком такого метода является возникновение отрицательных чисел и то, что наблюдения перестают быть независимыми.

6. Понижение размерности данных. При больших объемах данных, в том числе когда при наличии большого количества признаков, обработка данных ставится затруднительной, в связи с чем, аналитики стараются снизить размерность данных и не добавлять в набор данных признаки, которые являются малозначительными. Для того чтобы работать с признаками, которые несут полезную информацию, применяют различные методы, например, корреляционный или факторный анализ.

При отборе признаков с помощью коэффициентов корреляции необходимо оценить зависимость между входными признаками. Если между ними имеется сильная связь (коэффициент корреляции больше 0,9), то убирают из модели признак, имеющий наименьшую связь с зависимой переменной. Такой процесс продолжается до тех пор, пока все входные переменные будут независимыми (рис. 5).

Для снижения размерности был использован факторный анализ, который позволяет выявить признаки, объясняющие связь между собой. На выходе данного анализа получается набор данных, которые объясняют большую часть дисперсий наблюдений.

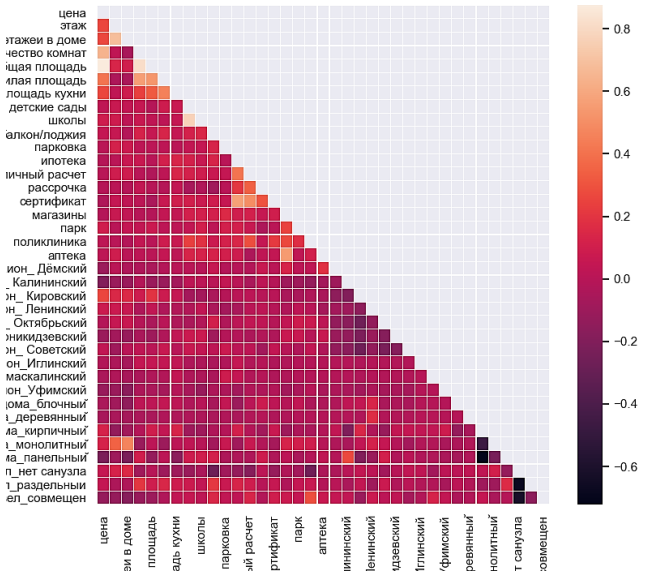


Рис. 5. Проверка корреляционной связи между входными признаками на примере вторичной недвижимости

Для проведения факторного анализа используется метод главных компонент, пошаговое описание метода представлено ниже.

Шаг 1. Вычисление среднего значения для каждого признака.

Шаг 2. Перенос начала координат в центр выборки.

Шаг 3. Вычисление ковариационной матрицы.

Шаг 4. Вычисление собственной композиции ковариационной матрицы.

Шаг 5. Получение списка собственных значений и списка собственных векторов

Шаг 6. Выбор необходимого количества (выбирается количество признаков, которое должно быть оставлено) наибольших собственных значений и соответствующих им собственных векторов.

Шаг 7. Проектирование выборки на выбранное направление.

После того как набор данных с допустимым качеством был получен, можно перейти к обработке данных.

Методы кластеризации [8–10]. В настоящее время существует множество алгоритмов кластеризации, применение которых на одних и тех же данных показывает различные результаты, а именно форму, размер или состав кластера. При выборе алгоритма нужно учитывать его специфику, например, склонность к созданию кластеров определенной или различной формы, склонности к определенному размеру при образовании групп, чувствительность к выбросам.

Методы разбиения на кластеры можно поделить на иерархические, которые не требуют определения количества кластеров, и неиерархические, в которых заранее определяется число кластеров. Иерархические делятся на агломеративные и дивизионные. Главное их отличие в том, что первые алгоритмы последовательно соединяют в одно древообъединение, а вторые – последовательно расщепляют группы. Иерархические алгоритмы являются наглядными, но предназначены только для небольших наборов данных.

Неиерархические разделяют объекты на кластеры итеративным способом, то есть объекты формируются в новые кластеры, пока не выполнится одно из условий остановки.

В исследовании были рассмотрены следующие методы: k-means, DBSCAN, самоорганизующиеся карты Кохонена, агломеративно иерархический метод. K-means относится к иерархическим алгоритмам кластеризации, в его основу положена следующая идея: данные произвольно разбиваются на кластеры, после чего итеративно пересчитывается центр масс для каждого кластера, полученного на предыдущем шаге, затем векторы разбиваются на кластеры вновь в соответствии с тем, какой из новых центров оказался ближе по выбранной метрике, алгоритм заканчивается, если центр масс не изменился по сравнению с предыдущим шагом. Алгоритм является простым для реализации, обеспечивает высокую скорость выполнения и хорошее качество кластеризации. Однако, алгоритм имеет ряд недостатков: количество кластеров является параметром алгоритма, высокая чувствительность к «выбросам» и начальному выбору центров тяжести, возможность сходимости к локальному оптимуму.

Алгоритм DBSCAN – плотностной алгоритм для кластеризации пространственных данных с присутствием шума. В данном алгоритме нет центроидов, кластеры формируются путем связывания соседних точек друг с другом. Алгоритм делает только один проход через данные, то есть если точке был назначен кластер, то он уже не будет изменен.

Достоинства данного метода: автоматическое выявление выбросов, устойчивость к шуму, алгоритм сам определяет количество кластеров, которые могут иметь произвольную форму (может отличать кластеры, которые нельзя разделить прямой линией), возможность масштабирования больших данных. Недостатки метода: не работает при различных плотностях кластера, в худшем случае алгоритмическая сложность равна $O(n^2)$, качество DBSCAN зависит от измерения расстояния.

Самоорганизующаяся карта Кохонена – нейронная сеть с обучением без учителя, выполняющая задачу визуализации и кластеризации. Сеть Кохонена имеет два слоя: входной и вы-

ходной, составленный из радиальных нейронов упорядоченной структуры. Цель данного метода: выявить скрытые закономерности в данных, понижая размерность чаще всего в двумерное пространство, удобное для представления и для интерпретации результатов.

Достоинствами метода являются устойчивость к зашумленным данным, быстрое и неуправляемое обучение, возможность упрощения многомерных входных данных с помощью визуализации. К недостаткам метода относится необходимость задавать число кластеров, зависимость результата от начальных заданных установок.

Агломеративно иерархический алгоритм группирует схожие объекты, начиная с маленьких кластеров и постепенно объединяя их в одну большую группу. Этот метод применяется для задач с близко расположенными кластерами.

Достоинством метода является автоматическое определение количества кластеров. Недостатки метода: если принято решение объединить два кластера, их уже невозможно разделить; алгоритм очень медленно работает на больших наборах данных, алгоритмическая сложность равна $O(n^2)$.

В исследовании рассматриваются различные выборки, а именно: вторичный и первичный рынок, аренда посуточная и аренда в месяц. Каждая выборка характеризуется своим набором признаков, поэтому было принято решение использовать все четыре алгоритма для сегментации объектов недвижимости на каждой выборке. Проведенный эксперимент позволил выбрать наилучший алгоритм для каждой выборки.

Для реализации задачи исследования недвижимости необходимо разделить ее на следующие подзадачи: добавление новой недвижимости в выбранную категорию данных, кластеризация выборки, отображение данных, вывод рекомендации по исследуемому объекту. Рассмотрим алгоритмы для каждой подзадачи.

Алгоритм подзадачи «Добавление новой недвижимости в выбранную категорию данных».

Шаг 1. Подготовка переменных для добавления информации о новой квартире.

Шаг 2. Обработка события для добавления данных.

Шаг 2.1. Проверка отсутствия ошибок при вводе признаков недвижимости.

Шаг 2.2. Если не были допущены ошибки при вводе информации, признаки добавляются в выбранный набор данных. Иначе выводится ошибка, описанная в руководстве пользователя (раздел «Сообщения»).

Алгоритм подзадачи «Кластеризация выборки».

Шаг 1. Открытие файла с данными в зависимости от выбранной категории недвижимости.

Шаг 2. Предобработка информации.

Шаг 3. Применение метода кластеризации в соответствии с выбором области недвижимости.

Алгоритм подзадачи «Отображение данных».

Шаг 1. Построение графического отображения кластеров по полученным меткам данных и сохранение переменной.

Шаг 2. Отображение полученной переменной в выделенном для него элементе интерфейса «канва».

Алгоритм подзадачи «Вывод рекомендации по исследуемому объекту».

Шаг 1. Определение принадлежности исследуемой квартиры к кластеру.

Шаг 2. Получение списка всех точек, принадлежащих этой группе.

Шаг 3. Вычисление значений максимальной, минимальной и средней цены данного кластера, вывод рекомендации при условии отсутствия признака, влияющего на стоимость объекта.

ПРОГРАММНОЕ РЕШЕНИЕ И ПРОВЕДЕНИЕ ЭКСПЕРИМЕНТА

При описании структуры программного решения следует отметить, что по одному адресу находятся четыре набора данных для анализа: первичная и вторичная выборка, данные о посуточной и месячной аренде. При кластеризации функция считывает, какой из этих файлов открыть для дальнейшей работы.

Класс Window отвечает за визуальное представление системы ввода/вывода. В нем содержится область для ввода данных, графическое представление сегментов недвижимости, характеристика кластера и таблица с похожими квартирами. Но для получения этих результатов внутри класса Window создается экземпляр класса Data_analysis, который, в свою очередь, получает информацию о выбранном объекте недвижимости и требования к квартире (рис. 6).

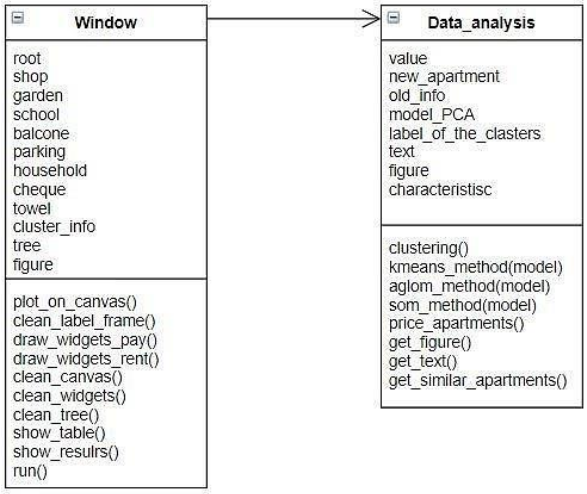


Рис. 6. Структура программного решения

Для реализации программного решения выбран графический инструмент tkinter для языка программирования Python.

Данный пакет работает с библиотекой Tk, и является многоплатформенным со свободно распространяемой лицензией, обладает простым и гибким синтаксисом, имеет удобные виджеты, сопровождается обширным руководством. Пример реализации программного решения представлен на рис. 7.

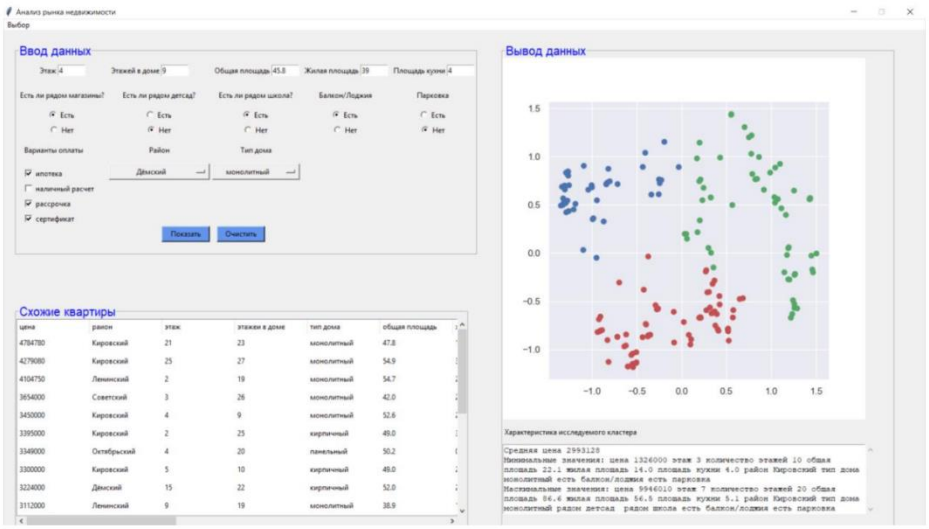


Рис. 7. Пример работы программного решения

Для выполнения эксперимента использовался класс ЭВМ со следующими характеристиками: ОС Windows 10x64; процессор AMDA6; 4 ГБ ОЗУ9-220.

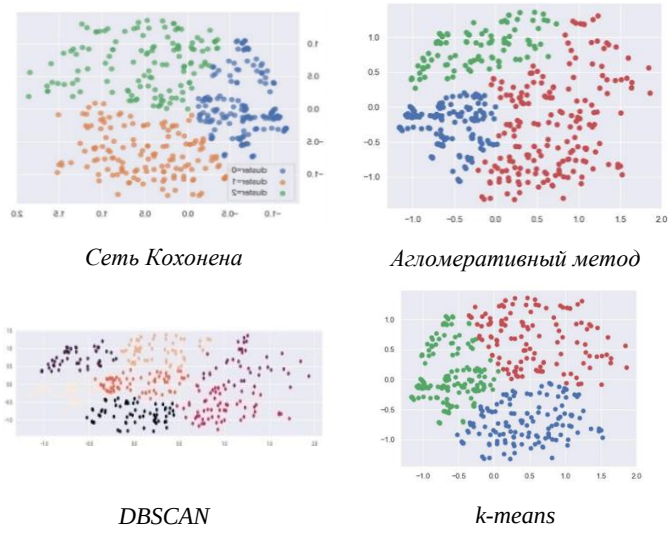
Целью эксперимента является выбор наилучшего алгоритма кластеризации по следующим параметрам: время выполнения алгоритма, качество разбиения объектов. В табл. 1 представлен наилучший и наихудший случай для каждой выборки.

Таблица 1

Результаты вычислительного эксперимента

	k-means	*Агломеративный метод*	*DBSCAN*	*Сеть Кохонена*
Вторичный рынок				
Время выполнения (секунды)	8,88	0,52	0,28	0,63
Силуэт	0,41	0,37	0,17	0,40
Первичный рынок				
Время выполнения (секунды)	4,68	0,38	0,39	0,48
Силуэт	0,51	0,44	0,43	0,51
Посуточная аренда				
Время выполнения (секунды)	2,54	0,31	0,29	0,44
Силуэт	0,5	0,47	0,35	0,49
Аренда в месяц				
Время выполнения (секунды)	2,07	0,35	0,31	0,43
Силуэт	0,49	0,48	0,33	0,48

В выборке вторичного рынка лучший результат достигнут нейронными сетями Кохонена, так как время выполнения намного меньше, чем у k-means, а значение силуэта примерно одинаковое. Наихудший результат получился с использованием алгоритма DBSCAN, так как кластеры пересекаются и выделено слишком много групп (рис. 8).



Сеть Кохонена *Агломеративный метод*

DBSCAN *k-means*

Рис. 8. Сегментация для вторичного жилья

В выборке первичного рынка был выбран алгоритм кластеризации k-means, так как по сравнению с другими методами кластеры распределены более равномерно. Алгоритм DBSCAN в данном случае не подходит из-за того, что один из кластеров собрал большую часть объектов, то есть пропорции кластеров резко различаются (рис. 9).

Для выборки аренды посуточно подходит агломеративный алгоритм кластеризации, так кластеры четко разделены и время выполнения допустимое.

Алгоритм k-means в данном случае не подойдет из-за длительного выполнения и смешивания кластеров (рис. 10).

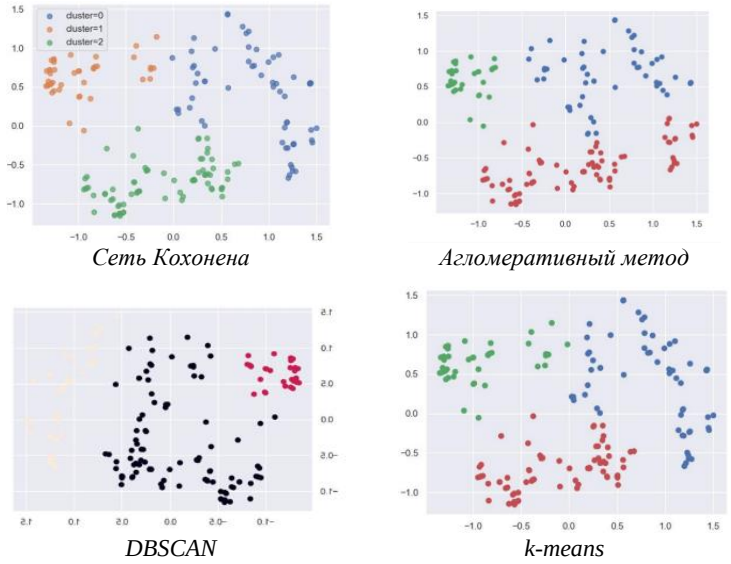


Рис. 9. Сегментация для первичного жилья

Для выборки аренды в месяц выбран агломеративный алгоритм кластеризации из- за приемлемого времени выполнения и качества кластеров. Метод k-means в данном случае не подошел, так как долго выполняется по сравнению с другими, и кластеры пересекаются (рис. 11).

Таким образом, для каждого набора данных осуществлен выбор наиболее эффективного метода.

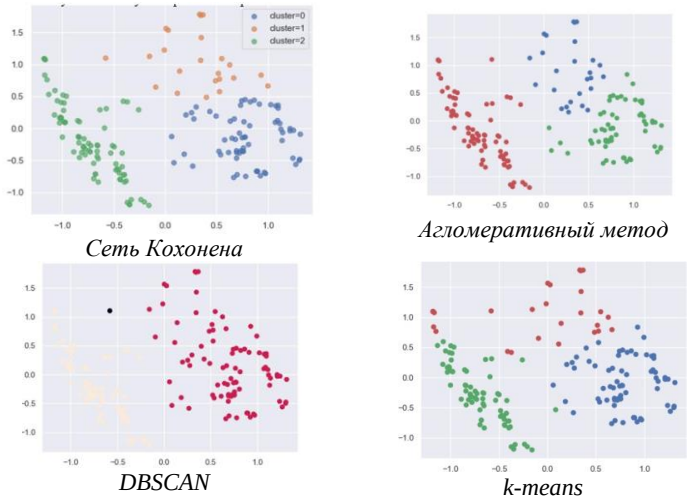


Рис. 10. Сегментация для набора данных «аренда посуточно»

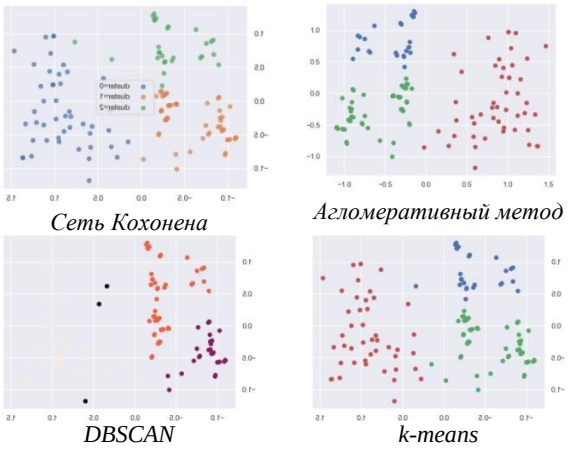


Рис. 11. Сегментация для набора данных «аренда в месяц»

ЗАКЛЮЧЕНИЕ

В статье рассматривается актуальная задача сегментирования на примере рынка недвижимости. Для ее решения предложен подход, включающий сбор информации и ее анализ, базовым компонентом которого является задача сегментирования.

Для сбора информации был выбран гибридный метод с использованием библиотеки lxml/html для языка Python. Выбор обоснован тем, что данная библиотека проста в использовании, обеспечивает преобразование HTML/XML кода в тип данных Python, упрощая работу с данными.

Анализ данных включает в себя этапы предобработки, сегментирования (кластеризации) и графического отображения кластеров. Для сегментации вторичного жилья были использованы самоорганизующиеся карты Кохонена, для первичного жилья – k-means, данные алгоритмы показали хорошие результаты при проведении экспериментов.

Созданное программное решение для сегментации рынка недвижимости позволяет графически представить сегменты объектов недвижимости, узнать характеристики объектов каждого сегмента, а также вывести пользователю сведения об объекте недвижимости, удовлетворяющем его требованиям.

Результаты исследования, связанные со сбором данных на основе гибридного подхода, их структурированием и использованием для проведения анализа, приведенные в статье, получены в рамках государственного задания № FEUE-2020-0007. Вопросы, отражающие структуру разработанного программного решения и визуализации результатов, получены в рамках выполнения гранта РФФИ 19-07-00709.

СПИСОК ЛИТЕРАТУРЫ

1. **Stoken D.** The Great Cycle: Predicting and Profiting from Crowd Behavior, the Kondratieff Wave and Long-Term Cycles. Revised Edition. Chicago, IL: Probus Publishing Company, 1993. 225 p. **[** D. Stoken, *The Great Cycle: Predicting and Profiting from Crowd Behavior, the Kondratieff Wave and Long-Term Cycles. Revised Edition.* Chicago, IL: Probus Publishing Company, 1993. **]**

2. **Управление** недвижимостью: учебник и практикум для академического бакалавриата / С. Н. Максимов [и др.]. М.: Юрайт, 2019. 416 с. **[** S. N. Maksimov, *et al.*, *Real estate management: textbook and workshop for bachelors,* (in Russian). Moscow: Yurait, 2019. **]**

3. **Стерник Г. М., Стерник С. Г.** Методология прогнозирования российского рынка недвижимости. Часть 1. Основные допущения, ограничения и рабочие гипотезы // Механизация строительства. 2013. № 8 (830). С. 53–63. **[** G. M. Sternik, S. G. Sternik, "Methodology for forecasting the Russian real estate market. Part 1. Basic assumptions, limitations and working hypotheses", (in Russian), in *Mekhanizaciya stroitelstva,* no. 8 (830), pp. 53-63, 2013. **]**

4. **Стерник Г. М., Стерник С. Г.** Анализ рынка недвижимости для профессионалов. Москва: Экономика, 2009. 608 с. **[** G. M. Sternik, S. G. Sternik, *Real estate market analysis for professionals,* (in Russian). Moscow: Economika, 2009. **]**

5. **Intelligent** Information Support for Decision Making in Maintenance and Equipment Repair Management / N. I. Yusupova, et al. // Proceedings of XXI International Conference Complex Systems: Control and Modeling Problems (CSCMP 2019), 2019. Pp. 192-197. DOI: 10.1109/CSCMP45713.2019.8976713. **[** N. I. Yusupova, *et al.,* "Intelligent Information Support for Decision Making in Maintenance and Equipment Repair Management", in *Proceedings of XXI International Conference Complex Systems: Control and Modeling Problems (CSCMP 2019),* pp. 192-197, 2019. **]**

6. **Knowledge** Identification by Structured Data for Decision Making in Project Teams / N. I. Yusupova, et al. // Proceedings of the 8th Scientific Conference on Information Technologies for Intelligent Decision Making Support (ITIDS 2020). Series Advances in Intelligent Systems Research. 2020. Vol. 174. Pp. 385-390. **[** Y. N. I.usupova, *et al.,* "Knowledge Identification by Structured Data for Decision Making in Project Teams", in *Proceedings of the 8th Scientific Conference on Information Technologies for Intelligent Decision Making Support (ITIDS 2020). Advances in Intelligent Systems Research,* vol. 174, pp. 385-390, 2020. **]**

7. **Юсупова Н. И., Сметанина О. Н., Гаянова М. М.** Технологии искусственного интеллекта и машинного обучения в задачах семантического представления и анализа данных: монография. М.: Инновационное машиностроение, 2020. 242 с. **[** N. I. Yusupova, O. N. Smetanina, M. M. Gayanova, *Technologies of artificial intelligence and machine learning in tasks of semantic representation and data analysis,* (in Russian). Moscow: Innovative machine building, 2020. **]**

8. **Грас Джоэл.** Data Science. Наука о данных с нуля. Санкт-Перербург: БХВ-Петербург, 2020. 336 с. **[** Joel Grus, *Data Science from Scratch,* (in Russian). Sankt-Peterburg: BXV-Peterburg, 2020. **]**

9. **Плас Дж. Вандер.** Python для сложных задач. Наука о данных и машинное обучение. СПб.: Питер, 2018. 576 с. **[** Plas Jake Vander, *Python Data Science Handbook. Essential Tools for Working with Data,* (in Russian). St. Petersburg: Piter, 2018. **]**

10. **Силен Д., Али М., Мейсман А.** Основы Data Science и Big Data. Python и наука о данных. СПб.: Питер, 2017, 336 с. **[** D. Cielen, A. Meysman, M. Ali, *Introducing Data Science. Big Data, Machine Learning, and More, Using Python Tools,* (in Russian). St. Petersburg: Piter, 2017. **]**

ОБ АВТОРАХ

ЯКУПОВА Айсылу Вазировна, стажер-консультант Регионального центра IBS в Уфе. Дипл. бакалавра по направлению подготовки «Математическое обеспечение и администрирование информационных систем» (УГАТУ, 2020). Иссл. в обл. поддержки принятия решений при управлении сложными объектами.

СМЕТАНИНА Ольга Николаевна, проф. каф. выч. мат. и киберн. Дипл. инж. по автоматиз. проц. обр. и выд. информации (УАИ, 1985). Д-р техн. наук (УГАТУ, 2012). Иссл. в обл. поддержки принятия решений при управлении слож-ными объектами.

САЗОНОВА (РАССАДНИКОВА) Екатерина Юрьевна, доц. каф. выч. математики и кибернетики. Дипл. экон.-мат. (УГАТУ, 2011). Иссл. в обл. поддержки принятия решений при управлении сложными объектами.

METADATA

Title: Software solution of segmentation problem based on intelligent technologies.

Authors: A. V. Yakupova[1], O. N. Smetanina[2], E. Yu. Sazonova[3]

Affiliation:

[1] IBS Regional Center in Ufa, Russia.

[2, 3] Ufa State Aviation Technical University (UGATU), Russia.

Email: [1]ajsylu.yakupowa@yandex.ru, [2]smoljushka@mail.ru, [3] ekaterina_rassadnikova@mail.ru

Language: Russian.

Source: Vestnik UGATU (scientific journal of Ufa State Aviation Technical University), vol. 25, no. 3 (93), pp. 132-144, 2021. ISSN 2225-2789 (Online), ISSN 1992-6502 (Print).

Abstract: In this article the issues of solving the segmentation problem objects are discussed, taking into account the significant characteristics, which, on the one hand, allows you to narrow the search for solutions, on the other hand, when analyzing data and interpreting the results of analysis, including segmentation to formalize recommendations. The formulation of the segmentation problem and the solution structure are given, as well as the choice of methods for performing segmentation and the development of algorithms. Particular attention is paid to the problem of data collection and their structuring for solving the problem, the rationale for using a hybrid method for collecting and the choice of technologies for the implementation of data collection. The authors propose to develop the software solution on the basis of formalized recommendations.

Key words: segmentation task; intelligent technologies; machine learning methods; data structuring; hybrid data collection method.

About authors:

YAKUPOVA, Aisylu Vazirovna, trainee consultant at IBS Regional Center in Ufa, bachelor of training “Mathematical support and administration of information systems” (USATU, 2020). Research works in the field of decision support in management of complex objects.

SMETANINA, Olga Nikolayevna, Prof., specialist in automated processing and delivery of information (UAI, 1985). Cand. of Tech. Sci. (UGATU, 1999), Dr. of Tech. Sci. (UGATU, 2012). Research works in the field of decision support in management of complex objects.

SAZONOVA (RASSADNIKOVA), Ekaterina Yurevna, Assoc. Prof. Dept. of Computational Mathematics and Cybernetics. Dipl. Economist-mathematician (USATU, 2011). Research works in the field of decision support in management of complex objects.