МИНОБРНАУКИ РОССИИ

Федеральное государственное бюджетное образовательное учреждение
высшего образования
**«САРАТОВСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
ИМЕНИ Н.Г. ЧЕРНЫШЕВСКОГО»**

Кафедра социальной информатики

КЛАСТЕРНЫЙ АНАЛИЗ КАК ТЕХНОЛОГИЯ СТРУКТУРИРОВАНИЯ СОЦИОЛОГИЧЕСКИХ ДАННЫХ

(автореферат бакалаврской работы)

студента 4 курса 451 группы
направления 09.03.03 - Прикладная информатика
профиль Прикладная информатика в социологии
Социологического факультета
Терентьева Артема Вячеславовича

Научный руководитель
профессор, кандидат социологических наук _____ С.В. Ситникова
подпись, дата

Зав. кафедрой
кандидат социологических наук, доцент _____ И.Г. Малинский
подпись, дата

Саратов 2021

ВВЕДЕНИЕ

Актуальность проблемы. В настоящее время практически не существует области интересов человека, в которой не существовало бы в той или иной степени потребности в обработке данных. В силу этого, задача выбора оптимального метода обработки данных зачастую является очень важной, особенно в тех случаях, когда необходима экспресс-оценка данных, приходящих из различных источников.

Любые методы обработки данных так или иначе используются для структурирования и анализа существующей информации. Задач по анализу информации много, однако в этой статье рассмотрены методы, которые эффективно работают для решения задач по структурированию данных с большим количеством разнородных параметров. Часто случается, что формальные методы анализа позволяют получить неожиданные новые знания.[1]

Существуют специальные математические методы, которые называются методами многомерной классификации, или методами кластерного анализа. Многомерная классификация представляет собой разбиение объектов на классы, в которой каждый объект может быть отнесен к нескольким классам, в зависимости от числа критериев разбиения. Кластерный анализ – это группа методов, используемых для классификации объектов или событий в относительно однородные группы, которые называют кластерами.

Основной целью в кластерном анализе является выделение сравнительно небольшого числа групп объектов, как можно более схожих между собой внутри группы, и как можно более отличающихся в разных группах. В информационных системах используется при решении задач классификации и обнаружения закономерностей в данных: при работе с базами данных, анализе интернет-документов, сегментации изображений и т.д. А также кластерный анализ широко используется во многих приложениях, включая исследования рынка, распознавание образов, анализ данных и обработку изображений. В бизнесе кластеризация может помочь маркетологам обнаружить отдельные группы в

[1] Рубаков С. В. Современные методы анализа данных. М. 2008. – 165с.

своих клиентских базах и охарактеризовать группы клиентов на основе моделей покупок.

В настоящее время разработано достаточно большое число алгоритмов кластерного анализа. Но универсальность использования вызвала появление большого количества несовместимых подходов, терминов и методов, которые затрудняют однозначное применение и непротиворечивую интерпретацию данного математического метода. не смотря на всю проделанную работу, единой системы классификации кластерных процедур на сегодняшний день не существует. Зачастую такие системы создаются отдельно в каждой отрасли применения кластерного анализа.

Степень научной разработанности данной проблемы.

Первые работы, посвященные описанию методам кластерного анализа, относятся к концу 1930-х годов. Активный интерес данной теме пришелся на период 60-80 гг.

Толчком для разработки многих кластерных методов послужила книга «Начала численной таксономии», опубликованная в 1963г. Робертом Сокэлом и Петером Снитом.

В настоящее время существует множество современных научных статей, посвященных специфике и сравнению применения кластеризации в различных сферах применения.

Поскольку анализ проводился на тему значимости видеоигр в повседневной жизни подростков, следует сказать, что эта проблема изучалась Кудиновым В. В. и Шишкиной К.И. в статье «Зависимость от компьютерных видеоигр у младших школьников»[2], а так же в научной статье «Влияние видеоигр на геймеров: к проблеме определения трансформационного потенциала игровой активности» Кыштымовой И. М. и Тимофеева С. Б.[3] и во многих других статьях.

[2] Кудинов В. В. Шишкина К.И. Зависимость от компьютерных видеоигр у младших школьников. СПб. 2019– с. 433-436.

[3] Кыштымовой И. М. и Тимофеева С. Б. Влияние видеоигр на геймеров: к проблеме определения трансформационного потенциала игровой активности. Иркутск. 2019 – с. 3.

Цель: Выявить специфику применения кластерного анализа как технологии структурирования социологических данных.

Задачи: 1. Представить и описать теоретические основания кластеризации статистических данных.

2. Рассмотреть особенности применения кластерного анализа в социологии.

3. Определить основные индикаторы значимости видеоигр в жизни современных подростков.

4. На основе кластерного анализа представить модель отношения подростков к видеоиграм.

Объектом данной работы – кластерный анализ, как технология структурирования социологических данных.

Предметом данной работы – специфика применения кластерного анализа в изучении роли видеоигр в повседневной жизни современных подростков.

Эмпирической базой выпускной квалификационной работы являются результаты авторского исследования на тему «Значимость видеоигр на повседневную жизнь подростков» (N=231 человек);

Теоретическая значимость исследования заключается в возможности использования основных положений и выводов данной работы для дальнейшего изучения выбранной проблематики.

Структура выпускной квалификационной работы представлена введением, тремя разделами, заключением и списком использованных источников.

ОСНОВНОЕ СОДЕРЖАНИЕ РАБОТЫ

В первом разделе «Теоретико-методологическое основание кластерного анализа» описывается специфика применения кластеризации. А также достоинства и недостатки популярных методов и мер расстояния.

Встречаются задачи, когда есть некоторая общая выборка наблюдений, в который предполагается наличие некоторой внутренней структуры. И как раз-таки методы кластерного анализа позволяют выделять группы схожих объектов внутри общей выборки. Говоря о выделении схожих объектов, прежде всего предполагается, что объекты, находящиеся в одной группе должны обладать схожими характеристиками. Поэтому появляется необходимость введения некоторой меры сходства для объектов.

Мера сходства должна основываться на каких-то признаках, которыми каждый объект описывается. Всего их можно выделить три: это меры расстояния, коэффициенты корреляции и коэффициенты ассоциативности. Каждая из этих мер имеет свои плюсы и минусы, области применения, и их все необходимо знать, потому что в зависимости от задачи нужно использовать различные меры.

Меры расстояния отталкиваются от понятия метрики, и при этом подходе объекты представляют собой точки в k-мерном пространстве, где размерность определяется количеством переменных, используемых для описания объектов.

Существует четыре наиболее распространенных расстояния, и первое из них — это евклидово расстояние, квадрат евклидова расстояния, манхэттенское расстояние и расстояние Чебышёва.

Существует так же и меры сходства, основанные на корреляции между наблюдениями. К ним чаще всего относят коэффициент корреляции Пирсона и расстояние Махаланобиса.

Во время работы с бинарными признаками, в качестве мер сходства считают так называемые коэффициенты ассоциативности. Самый популярный из них — это так называемый φ-коэффициент.

Большинство современных кластерных методов относятся к семейству иерархических. Иерархический кластерный анализ – это совокупность алгоритмов упорядочивания данных, визуализация которых обеспечивается с помощью графов. Работа большинства таких методов основана на вычислении матрицы сходства, которая содержит меры расстояний.

Среди иерархических методов присутствует центроидный, межгрупповой связи и метод Уорда. А также методы ближайшего и удаленного соседа.

Так же существуют и другие популярные эффективные методы кластеризации: k-means, k-median, С-средних, EM-алгоритм и алгоритм кластеризации FOREL.

Во втором разделе «Практика применения кластерного анализа в социально-гуманитарных науках» описывается применение кластерных методов анализа с подробной интерпретацией алгоритма.

Первый пример применения кластерного анализа описывает исследование для оценки социально-экономического уровня развития регионов.

Кластерный анализ был осуществлен для всех областей ЦФО за 4 года с 2008 г. по 2011 г. В данной работе в качестве объектов исследования выступили 17 регионов, входящих в ЦФО.

Прежде чем осуществить кластерный анализ, исследователями были выделены наиболее значимые характеристики уровня социально-экономического развития необходимые для оценки социально-экономического положения субъектов ЦФО. Далее был проведен поэтапный кластерный анализ на базе 10 статистических показателей, которые наиболее полно характеризуют соответствующие условия развития регионов. В качестве результирующей величины был взят валовой региональный продукт (ВРП) на душу населения.

Области решили разбить на три кластера в связи с тем, что результат исследования должен был показать слабые, средние и сильные подгруппы. Каждой группе дали своё условное название в зависимости от характерных черт. Первую группу назвали «лидерами». Вторую, самую объемную, – «последователями», а третью «аутсайдерами».

Проведенная кластеризация позволила исследователям провести дальнейший расчет интегрального показателя социально-экономического развития области, входящей в федеральный округ.

Проведение сравнительной оценки социально-экономического развития областей ЦФО с применением метода кластерного анализа, позволило выделить группы областей округа со сходным сочетанием значений признаков, а также определить место и роль каждого из них в экономике, что имеет большое значение для разработки важнейших для экономики целевых программ, направления инвестиций, государственной поддержки нуждающихся регионов.

Второй пример применения катерного анализа описывает исследование, направленное на повышение эффективности работы медицинских учреждений.

Медицинские учреждения располагают большими массивами данных, однако традиционный подход документирования процессов не позволяет сформировать полную картину всех траекторий пациентов и использовать накопленные данные для решения оптимизационных задач. Кроме того, разнообразная природа заболеваний отражается в высокой вариативности маршрутов.

Клинический путь пациента может включать в себя цепочку соответствующих профилю медицинского учреждения событий: первичную запись на прием к терапевту, лабораторные исследования, получение консультации узкопрофильного специалиста и другие.

Разработанные клинические пути интегрировались в электронный документооборот медицинских учреждений. Однако стремительный рост массивов доступных данных, включая изображения в цифровом виде, выявил необходимость автоматического определения клинического пути пациента на основе этих данных. Решением для автоматического выявления персонального клинического пути стали технологии интеллектуального анализа процессов, интеллектуального анализа данных, алгоритмов машинного обучения и другие.

Для поиска модели, в которой находится кластер с максимальным коэффициентом силуэта, сравнивались метод Уорда и k-medoids,

напоминающий метод k-means. В соответствии с проведенным анализом значений коэффициента был выбран метод Уорда и были выявлены тенденции полученных групп.

Далее была построена карта процессов исходного датасета, однако без предварительного разделения маршрутов на кластеры сложно интерпретировать полученные клинические пути. После определения оптимального количества кластеров были отдельно сформированы процессные карты для полученных групп.

Следующим этапом была нечеткая кластеризация исходных данных методами тематического моделирования, при которой путь пациента может относиться к нескольким шаблонам с различными вероятностями.

Полученные результаты позволили провести предварительную оценку клинических путей пациентов любого журнала событий, определить узкие места системы и визуализировать процессные карты деятельности медицинского учреждения. Метод нечеткой кластеризации позволил добавить иерархическое представление маршрутов пациентов, отображая ресурсы медицинских учреждений. Выделенные кластеры будут отправной точкой для улучшения прогноза потока прикрепленного контингента, а также для формирования рекомендаций по ресурсному оснащению больниц при развитии сервисов.

Третий раздел «Кластерная модель значимости видеоигр в повседневной жизни подростков» построен на применении кластерного анализа для авторского социологического исследования, отражающего влияние видеоигр на повседневную жизнь подростков.

Полностью проанализировать сами видеоигры как явление довольно сложно из-за того, что каждый день появляются новые игры и игровые сообщества. Вектор развития игровой индустрии постоянно меняется в связи с появлением новых игровых жанров и сегодня становится точкой бифуркации для определения роли и места компьютерных игр в истории человечества. В современном обществе пока нет людей, которые выросли исключительно на компьютерных играх, но образовательный и пропагандистский потенциал

компьютерных игр и их влияние на внутренний мир человека должны оцениваться прямо сейчас. Поэтому так важно понимать природу компьютерных игр и тенденции, которые в настоящее время складываются в современном обществе, что делает, на наш взгляд, эту проблему актуальной.

В текущем исследовании были применены такие методы кластеризации, как метод Уорда и k-средних. В качестве мер сходства использовалось Манхэттенское расстояние и квадрат Евклидова расстояния.

Первая кластеризация позволила определить, как отношение родителей к увлечению видеоиграми и материальное положение влияет на успехи подростков в школьной программе.

Как оказалось, подростки, находящиеся в семьях с невысоким уровнем дохода и не получающие от родителей никакой реакции на увлечение видеоиграми, оказываются наименее успешными. А подростки, находящиеся в семьях с низким уровнем дохода и активным влиянием родителей, оказываются наиболее успешными в учебе.

Вторая кластеризация позволила объединить подростков в четыре группы, определяющие игровые предпочтения и в какой-то степени, указывающие на темперамент респондента.

Получилось выделить следующие группы: в первый кластер входят подростки, предпочитающие те игры, которые требуют от пользователя *вдумчивый и усидчивый настрой*. Ко второму кластеру принадлежат подростки, предпочитающие максимально использовать свою *концентрацию и внимание*. К третьему кластеру относятся подростки, предпочитающие испытывать максимум *эмоций* от времяпровождения. И наконец, четвертый кластер, включает в себя подростков, желающих удовлетворить свой *соревновательный дух* в файтинге или спортивной игре.

Далее была рассмотрена корреляция кластеров с вопросом об увлечениях респондентов. Как оказалось, в первом кластере отличаются такие увлечения, как спорт, наука, творчество и фильмы. Во втором кластере наиболее популярны творческие увлечения и литература различного жанра. Третий кластер

отличается относительно большим количеством подростков, в чьи интересы входят практически все имеющиеся в списке увлечения, за исключением коллекционирования. А четвертый кластер отличается самой большой долей по выбору спортивных увлечений, так же в нем наибольшее часто встречаются киноманы и коллекционеры среди подростков, играющих в видеоигры.

ЗАКЛЮЧЕНИЕ

При подготовке дипломной работы автором была определена цель – выявить специфику применения кластерного анализа как технологии структурирования социологических данных. Для реализации данной цели им были описаны наиболее популярные методы кластеризации и определяемые меры сходства. Так же были описаны случаи, для которых свойственно применять тот или иной метод. Для определения выбора меры сходства были описаны подходящие типы данных и их области применения.

В качестве примеров анализа данных методом кластеризации, автором были описаны и изучены исследовательские работы в сфере социально-экономического развития, где был проведен анализ динамики экономической активности регионов и в сфере медицины, где проведенный анализ позволил провести предварительную оценку клинических путей пациентов.

В работе автора было проиллюстрировано то, как кластерный анализ позволяет производить разбиение объектов не по одному параметру, а по целому набору признаков. Кроме того, автор показал, что кластерный анализ в отличие от большинства математико-статистических методов не накладывает никаких ограничений на вид рассматриваемых объектов и позволяет рассматривать множество исходных данных в различном виде.

Кластерный анализ позволяет рассматривать достаточно большой объем информации и резко сокращать, сжимать большие массивы информации, делать их компактными и наглядными. А также позволяет не только выявлять тенденционные направления, но и определять их содержательные особенности. Кроме того, это позволяет делать прогнозы, экстраполируя данные на всю генеральную совокупность.