Интеллектуальный анализ данных

Online Analytical Processing – аналитическая обработка данных в реальном времени

OLAP-системы предоставляют аналитику средства проверки гипотез при анализе данных. При этом основной задачей аналитика является генерация гипотез. Он решает ее, основываясь на своих знаниях и опыте. Однако знания есть не только у человека, но и в накопленных данных, которые подвергаются анализу. Такие знания часто называют «скрытыми», т.к. они содержатся в гигабайтах и терабайтах информации, которые человек не в состоянии исследовать самостоятельно. В связи с этим существует высокая вероятность пропустить гипотезы, которые могут принести значительную выгоду.

Data Mining (Добыча данных) – исследование и обнаружение «машиной» (алгоритмами, средствами искусственного интеллекта) в сырых данных скрытых знаний, которые ранее не были известны, нетривиальны, практически полезны, доступны для интерпретации человеком.

Практическое применение Data Mining.

Интернет-торговля:

В системах электронного бизнеса, где особую важность имеют вопросы привлечения и удержания клиентов, технологии Data Mining часто применяются для построения рекомендательных систем интернет-магазинов и для решения проблемы персонализации посетителей Web-сайтов.

- анализ траекторий покупателей от посещения сайта до покупки товаров
- оценка эффективности обслуживания, анализ отказов в связи с отсутствием товаров
- связь товаров, которые интересны посетителям

Торговля

Для успешного продвижения товаров всегда важно знать, что и как продается, а также, кто является потребителем. Исчерпывающий ответ на первый вопрос дают такие средства Data Mining, как анализ рыночных корзин и сиквенциальный анализ. Зная связи между покупками и временные закономерности, можно оптимальным образом регулировать предложение. С другой стороны, маркетинг имеет возможность непосредственно управлять спросом, но для этого необходимо знать как можно больше о потребителях – целевой аудитории маркетинга. Data Mining позволяет решать задачи выделения групп потребителей со схожими стереотипами поведения, т. е. сегментировать рынок. Для этого можно применять такие технологии Data Mining, как кластеризацию и классификацию

- *анализ покупательской корзины*;
- *создание предсказательных моделей* и классификационных моделей покупателей и *покупаемых товаров;*
- создание профилей покупателей;
- оценка лояльности покупателей разных категорий лояльности;
- *исследование временных рядов* и временных зависимостей*, выделение сезонных факторов, оценка эффективности рекламных акций на большом диапазоне реальных данных.*

Телекоммуникации

Телекоммуникационный бизнес является одной из наиболее динамически развивающихся областей современной экономики. Возможно, поэтому традиционные проблемы, с которыми сталкивается в своей деятельности любая компания, здесь ощущаются особо остро.

Телекоммуникационные компании работают в условиях жесткой конкуренции, что проявляется в ежегодном оттоке около 25 % клиентов.

- *классификация клиентов на основе ключевых характеристик вызовов (частота, длительность и т.д.), частоты смс;*
- *выявление лояльности клиентов;*
- *определение мошенничества и др.*

Промышленное производство

Промышленное производство создает идеальные условия для применения технологий Data Mining. Причина – в самой природе технологического процесса, который должен быть воспроизводимым и контролируемым. Все отклонения в течение процесса, влияющие на качество выходного результата, также находятся в заранее известных пределах. Таким образом, создается статистическая стабильность, первостепенную важность которой отмечают в работах по классификации. Естественно, что в таких условиях использование Data Mining способно дать лучшие результаты, чем, к примеру, при прогнозировании ухода клиентов телекоммуникационных компаний.

Медицина

В медицинских и биологических исследованиях, равно как и в практической медицине, спектр решаемых задач настолько широк, что возможно использование любых методологий Data Mining. Примером может служить построение диагностической системы или исследование эффективности хирургического вмешательства.

Известно много экспертных систем для постановки медицинских диагнозов. Они построены главным образом на основе правил, описывающих сочетания различных симптомов отдельных заболеваний. С помощью таких правил узнают не только, чем болен пациент, но и как нужно его лечить. Правила помогают выбирать средства медикаментозного воздействия, определять показания/противопоказания, ориентироваться в лечебных процедурах, создавать условия наиболее эффективного лечения, предсказывать исходы назначенного курса лечения и т. п. Технологии Data Mining позволяют обнаруживать в медицинских данных шаблоны, составляющие основу указанных правил.

Банковское дело

Классическим примером использования Data Mining на практике является решение проблемы о возможной некредитоспособности клиентов банка.

Использование технологии Data Mining позволяет сократить число нарушений на 20–30 %.

Страховой бизнес

В страховании, так же как в банковском деле и маркетинге, возникает задача обработки больших объемов информации для определения типичных групп (профилей) клиентов. Эта информация используется для того, чтобы предлагать определенные услуги страхования с наименьшим для компании риском и, возможно, с пользой для клиента.

Другие области применения

Data Mining может применяться практически везде, где возникает задача автоматического анализа данных. В качестве примера приведем такие популярные направления, как анализ и последующая фильтрация спама, а также разработка так называемых виртуальных собеседников.

Процесс обнаружения знаний
Основные этапы анализа

Весь процесс можно разбить на следующие этапы:

1. понимание и формулировка задачи анализа;
2. подготовка данных для автоматизированного анализа (препроцессинг);
3. применение методов Data Mining и построение моделей;
4. проверка построенных моделей;
5. интерпретация моделей человеком.

На первом этапе выполняется осмысление поставленной задачи и уточнение целей, которые должны быть достигнуты методами Data Mining. Важно правильно сформулировать цели и выбрать необходимые для их достижения методы, т. к. от этого зависит дальнейшая эффективность всего процесса.

Второй этап состоит в приведении данных к форме, пригодной для применения конкретных методов Data Mining, вид преобразований, совершаемых над данными, во многом зависит от используемых методов, выбранных на предыдущем этапе.

Третий этап – это собственно применение методов Data Mining Сценарии этого применения могут быть самыми различными и включать сложную комбинацию разных методов, особенно если используемые методы позволяют проанализировать данные с разных точек зрения

Следующий этап – проверка построенных моделей. Очень простой и часто используемый способ заключается в том, что все имеющиеся данные, которые необходимо анализировать, разбиваются на две группы. Как правило, одна из них большего размера, другая – меньшего

Последний этап – интерпретация полученных моделей человеком в целях их использования для принятия решений, добавление получившихся правил и зависимостей в базы знаний и т.д. Этот этап часто подразумевает использование методов, находящихся на стыке технологии Data Mining и технологии экспертных систем.

Текущее состояние дел

•Точно знаем надо
• Примерно знаем почему
•Плохо знаем как

Данные

• Собираются не для анализа
• Собираются не всегда, когда можно
• Собираются некачественно

Проблемы

• Малая выборка
•Несоблюдение чистоты
•Недооценка динамики
•Недоверие к первым результатам

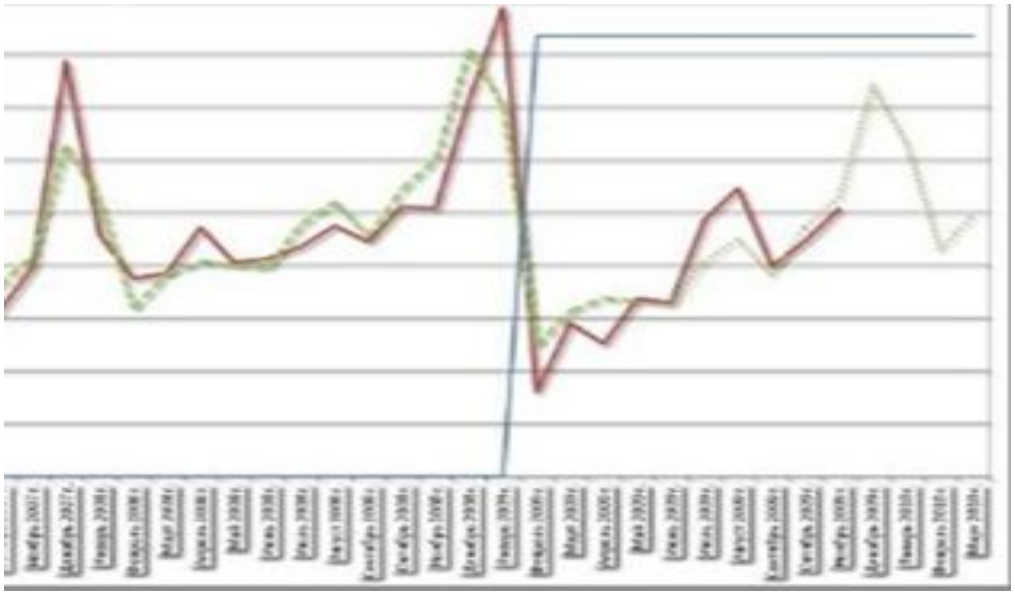
«Ручное» прогнозирование

Стратегия: выявить шаблоны «вручную»

Примеры (реальные случаи)
ошибки при вводе марки автомобиля:
14 (!)вариантов написания марки “Mercedes”.
DEU указано вместо DAEWOO в 6-ти анкетах,
Все заемщики рассчитались с кредитом.
Ошибочный вывод: наличие автомобиля марки DEU свидетельствует о высокой надежности клиента;

указана область проживания как БРЕСЦКАЯ (4 случая– все «плохие»). На практике выяснилось, что значимость региона не столь высока;
количество не столь очевидных примеров велико.
Доля строк хотя бы с одной ошибкой, опечаткой или пропуском может достигать 70%.

Клиенты приходят в разное время и их качественный состав меняется
Измерения производятся точно,
результаты тщательно регистрируются
Работают люди: ошибаются, пропускают, путают
Отбираются образцы в пропорциях, отражающих реальное положение дел
Есть сведения только о клиентах, получивших одобрение на выдачу кредита



Продажа стиральных машин

Продажа
майонеза

Классификация задач Data Mining

Методы DM помогают решить многие задачи, с которыми сталкивается аналитик. Из них основными являются: классификация, регрессия, поиск ассоциативных правил и кластеризация.

Задача классификации сводится к определению класса объекта по его характеристикам. Необходимо заметить, что в этой задаче множество классов, к которым может быть отнесен объект, заранее известно.

Задача регрессии, подобно задаче классификации, позволяет определить по известным характеристикам объекта значение некоторого его параметра. В отличие от задачи классификации значением параметра является не конечное множество классов, а множество действительных чисел.

При поиске ассоциативных правил целью является нахождение частых зависимостей (или ассоциаций) между объектами или событиями. Найденные зависимости представляются в виде правил и могут быть использованы как для лучшего понимания природы анализируемых данных, так и для предсказания появления событий.

Задача кластеризации заключается в поиске независимых групп (кластеров) и их характеристик во всем множестве анализируемых данных. Решение этой задачи помогает лучше понять данные. Кроме того, группировка однородных объектов позволяет сократить их число, а следовательно, и облегчить анализ.

Перечисленные задачи по назначению делятся на описательные и предсказательные.

Описательные (descriptive) задачи уделяют внимание улучшению понимания анализируемых данных. Ключевой момент в таких моделях – легкость и прозрачность результатов для восприятия человеком. Возможно, обнаруженные закономерности будут специфической чертой именно конкретных исследуемых данных и больше нигде не встретятся, но это все равно может быть полезно и потому должно быть известно. К такому виду задач относятся кластеризация и поиск ассоциативных правил.

Решение предсказательных (predictive) задач разбивается на два этапа.

На первом этапе на основании набора данных с известными результатами строится модель.

На втором этапе она используется для предсказания результатов на основании новых наборов данных. При этом, естественно, требуется, чтобы построенные модели работали максимально точно. К данному виду задач относят задачи классификации и регрессии. Сюда можно отнести и задачу поиска ассоциативных правил, если результаты ее решения могут быть использованы для предсказания появления некоторых событий.

Кластеризация

Задача кластеризации состоит в разделении исследуемого множества объектов на группы «похожих» объектов, называемых кластерами. Часто решение задачи разбиения множества элементов на кластеры называют кластерным анализом.

Кластеризация может применяться практически в любой области, где необходимо исследование экспериментальных или статистических данных. Рассмотрим пример из области маркетинга, в котором данная задача называется сегментацией.

Концептуально сегментирование основано на предпосылке, что все потребители – разные. У них разные потребности, разные требования к товару, они ведут себя по-разному: в процессе выбора товара, в процессе приобретения товара, в процессе использования товара, в процессе формирования реакции на товар. В связи с этим необходимо по-разному подходить к работе с потребителями: предлагать им различные по своим характеристикам товары, по-разному продвигать и продавать товары. Для того чтобы определить, чем отличаются потребители друг от друга и как эти отличия отражаются на требованиях к товару, и производится сегментирование потребителей.

Постановка задачи кластеризации

Кластеризация отличается от классификации тем, что для проведения анализа не требуется иметь выделенную целевую переменную, с этой точки зрения она относится к классу *unsupervised learning*. Эта задача решается на начальных этапах исследования, когда о данных мало что известно. Ее решение помогает лучше понять данные, и с этой точки зрения задача кластеризации является описательной задачей.

Для этапа кластеризации характерно отсутствие каких-либо различий как между переменными, так и между записями. Напротив, ищутся группы наиболее близких, похожих записей. Методы автоматического разбиения на кластеры редко используются сами по себе, просто для получения групп схожих объектов. Анализ только начинается с разбиения на кластеры. После определения кластеров используются другие методы, для того чтобы попытаться установить, а что означает такое разбиение на кластеры, чем оно вызвано.

Большое достоинство кластерного анализа в том, что он позволяет производить разбиение объектов не по одному параметру, а по целому набору признаков. Кроме того, кластерный анализ, в отличие от большинства математико-статистических методов, не накладывает никаких ограничений на вид рассматриваемых объектов и позволяет рассматривать множество исходных данных практически произвольной природы.

Формальная постановка задачи

Дано — набор данных со следующими свойствами:

- каждый экземпляр данных выражается четким числовым значением;
- класс для каждого конкретного экземпляра данных неизвестен.

Найти:

- способ сравнения данных между собой (меру сходства);
- способ кластеризации;
- разбиение данных по кластерам.

Формально задача кластеризации описывается следующим образом.

Дано множество объектов данных *I*, каждый из которых представлен набором атрибутов. Требуется построить множество кластеров *C* и отображение *F* множества *I* на множество *C*, т. е. $F: I \rightarrow C$. Отображение *F* задает модель данных, являющуюся решением задачи. Качество решения задачи определяется количеством верно классифицированных объектов данных.

Множество *I* определим следующим образом:
$I = \{i_1, i_2, \ldots, i_j, \ldots, i_n\}$,
где i_j — исследуемый объект.

Меры близости, основанные на расстояниях, используемые в алгоритмах кластеризации

Расстояния между объектами предполагают их представление в виде точек *m*-мерного пространства R^m. В этом случае могут быть использованы различные подходы к вычислению расстояний.

Рассмотренные ниже меры определяют расстояния между двумя точками, принадлежащими пространству входных переменных. Используются следующие обозначения:

$X_Q \subseteq R^m$ — множество данных, являющееся подмножеством *m*-мерного вещественного пространства;

$x_i = (x_{i1}, x_{i2}, \ldots, x_{im})$ — элементы множества данных;

$\bar{x} = \frac{1}{Q}\sum_{i=1}^{Q} x_i$ — среднее значение точек данных;

$S = \frac{1}{Q-1}\sum_{i=1}^{Q}\left(x_i - \bar{x}\right)\left(x_i - \bar{x}\right)'$ — ковариационная матрица (*m*×*n*).

Евклидово расстояние. Иногда может возникнуть желание возвести в квадрат стандартное евклидово расстояние, чтобы придать большие веса более отдаленным друг от друга объектам. Это расстояние вычисляется следующим образом:

$$d_2\left(x_i, x_j\right) = \sqrt{\sum_{t=1}^{m}\left(x_{it} - x_{jt}\right)^2}$$

Расстояние по Хеммингу. Это расстояние является просто средним разностей по координатам. В большинстве случаев данная мера расстояния приводит к таким же результатам, как и для обычного расстояния Евклида, однако для нее влияние отдельных больших разностей (выбросов) уменьшается (т. к. они не возводятся в квадрат). Расстояние по Хеммингу вычисляется по формуле

$$d_H\left(x_i, x_j\right) = \sum_{t=1}^{m}\left|x_{it} - x_{jt}\right|$$

Расстояние Чебышева. Это расстояние может оказаться полезным, когда желают определить два объекта как «различные», если они различаются по какой-либо одной координате (каким-либо одним измерением). Расстояние Чебышева вычисляется по формуле

$$d_{\infty}\left(x_i, x_j\right) = \max_{1 \le t \le m} \left| x_{it} - x_{jt} \right|$$

Расстояние Махаланобиса преодолевает этот недостаток, но данная мера расстояния плохо работает, если ковариационная матрица высчитывается на всем множестве входных данных. В то же время, будучи сосредоточенной на конкретном классе (группе данных), данная мера расстояния показывает хорошие результаты:

$$d_M\left(x_i, x_j\right) = \left(x_i - x_j\right) S^{-1} \left(x_i - x_j\right)^t$$

Пиковое расстояние предполагает независимость между случайными переменными, что говорит о расстоянии в ортогональном пространстве. Но в практических приложениях эти переменные не являются независимыми:

$$d_L\left(x_i, x_j\right) = \frac{1}{m}\sum_{t=1}^{m}\frac{\left|x_{it} - x_{jt}\right|}{x_{it} - x_{jt}}$$

Представление результатов

Результатом кластерного анализа является набор кластеров, содержащих элементы исходного множества. Кластерная модель должна описывать как сами кластеры, так и принадлежность каждого объекта к одному из них.

Для небольшого числа объектов, характеризующихся двумя переменными, результаты кластерного анализа изображают графически. Элементы представляются точками, кластеры разделяются прямыми, которые описываются линейными функциями.

Дивизимные алгоритмы

Дивизимные кластерные алгоритмы, в отличие от агломеративных, на первом шаге представляют все множество элементов *I* как единственный кластер. На каждом шаге алгоритма один из существующих кластеров рекурсивно делится на два дочерних. Таким образом итерационно образуются кластеры сверху вниз. Его применяют, когда необходимо разделить все множество объектов *I* на относительно небольшое количество кластеров.

Задача классификации и регрессии

При анализе часто требуется определить, к какому из известных классов относятся исследуемые объекты, т е. классифицировать их. Например, когда человек обращается в банк за предоставлением ему кредита, банковский служащий должен принять решение, кредитоспособен ли потенциальный клиент или нет. Очевидно, что такое решение принимается на основании данных об исследуемом объекте (в данном случае – человеке), его месте работы, размере заработной платы, возрасте, составе семьи и т.п. В результате анализа этой информации банковский служащий должен отнести человека к одному из двух известных классов «кредитоспособен» и «некредитоспособен».

Задача поиска ассоциативных правил

Поиск ассоциативных правил является одним из самых популярных приложений Data Mining. Суть задачи заключается в определении часто встречающихся наборов объектов в большом множестве таких наборов. Данная задача является частным случаем задачи классификации. Первоначально она решалась при анализе тенденций в поведении покупателей в супермаркетах. Анализу подвергались данные о совершаемых ими покупках, которые покупатели складывают в тележку (корзину). Это послужило причиной второго часто встречающегося названия – анализ рыночных корзин (Basket Analysis).

Задача поиска ассоциативных правил предполагает отыскание частых наборов в большом числе наборов данных.
В контексте анализы рыночной корзины это поиск наборов товаров, которые наиболее часто покупаются вместе.
В задаче не учитывался такой атрибут транзакции как время. Тем не менее, взаимосвязь событий во времени также представляет большой интерес. Основываясь на том, какие события чаще всего следуют за другими, можно заранее предсказывать их появление, что позволит принимать более правильные решения.

Отличие поиска ассоциативных правил от секвенциального анализа (анализа последовательностей) в том, что в первом случае ищется набор объектов в рамках одной транзакции, т.е. такие товары, которые чаще всего покупаются ВМЕСТЕ. В одно время, за одну транзакцию.
Во втором же случае ищутся не часто встречающиеся наборы, а часто встречающиеся последовательности.
Т.е. в какой последовательности покупаются товары или через какой промежуток времени после покупки товара "А", человек наиболее склонен купить товар "Б". Т. е. данные по одному и тому же клиенту, но взятые из разных транзакций.

Получаемые закономерности в действиях покупателей можно использовать для формирования более выгодного предложения, стимулирования продаж определённых товаров, управления запасами и т.п.

Секвенциальный анализ актуален и для телекоммуникационных компаний. Основная проблема, для решения которой он используется, - это анализ данных об авариях на различных узлах телекоммуникационной сети. Информация о последовательности совершения аварий может помочь в обнаружении неполадок и предупреждении новых аварий.

Введём некоторые обозначения и определения.

D - множество всех транзакций T, где каждая транзакция характеризуется уникальным идентификатором покупателя, временем транзакции и идентификатором объекта (id товара);
I - множество всех объектов (товаров) общим числом m;
s_i - набор, состоящий из элементов множества I;
S - последовательность, состоящая из различных наборов s_i;

Дальнейшие рассуждения строятся на том, что в любой случайно выбранный момент времени у покупателя не может быть более одной транзакции.
Шаблон последовательности - это последовательность наборов, которая часто встречается в транзакциях (в определённом порядке).
Последовательность $< a_1,a_2,...,a_n >$ **является входящей** в последовательность $< b_1,b_2,...,b_n >$, если существуют такие $i_1 < i_2 < i_n$, при которых $a_1 \subseteq b_{i_1}, a_2 \subseteq b_{i_2}, ..., a_n \subseteq b_{i_n}$

Например, последовательность <(3)(6,7,9)(7,9)> входит в <(2)(3)(6,7,8,9)(7)(7,9)>, поскольку

$$(3) \subseteq (3), (6,7,9) \subseteq (6,7,8,9), (7,9) \subseteq (7,9)$$

Поддержка последовательности - это отношение числа покупателей, в чьих транзакциях присутствует указанная последовательность к общему числу покупателей.
Также как и в задаче поиска ассоциативных правил применяется минимальная и максимальная поддержка. Минимальная поддержка позволяет исключить из рассмотрения последовательности, которые не являются частыми. Максимальная поддержка исключает очевидные закономерности в появлении последовательностей. Оба параметра задаются пользователем до начала работы алгоритма.

Алгоритм AprioriALL

Существует большое число разновидностей алгоритма Apriori, который изначально не учитывал временную составляющую в наборах данных.

Первым алгоритмом на основе Apriori, позволившим находить закономерности в последовательностях событий, стал предложенный в 1995 году (Argwal и Srikant) алгоритм AprioriALL.

Данный алгоритм, также как другие усовершенствования Apriori основывается на утверждении, что последовательность, входящая в часто встречающуюся последовательность, также является часто встречающейся.

Формат данных, с которыми работает алгоритм :

Это таблица транзакций с тремя атрибутами (id клиента, время транзакции, id товаров в наборе).

Работа алгоритма состоит из нескольких фаз.

Фаза сортировки заключается в перегруппировке записей в таблице транзакций. Сперва записи сортируются по уникальному ключу покупателя, а затем по времени внутри каждой группы.

Идентификатор покупателя	Время транзакции	Идентификаторы купленных товаров
1	Окт 23 08	30
1	Окт 28 08	90
2	Окт 18 08	10, 20
2	Окт 21 08	30
2	Окт 27 08	40, 60, 70
3	Окт 15 08	30, 50, 70
4	Окт 8 08	30
4	Окт 16 08	40, 70
4	Окт 25 08	90
5	Окт 20 08	90

Фаза отбора кандидатов - в исходном наборе данных производится поиск последовательностей в соответствии со значением минимальной поддержки. Предположим, что значение минимальной поддержки 40%.
Обратим внимание, что поддержка рассчитывается не из числа транзакций, в которые входит последовательность (в данном случае это есть набор), но из числа покупателей у которых во всех их транзакциях встречается данная последовательность.
В результате получим следующие последовательности.

Последовательности	**Идентификатор последовательности**
(30)	1
(40)	2
(70)	3
(40, 70)	4
(90)	5

Фаза трансформации. В ходе работы алгоритма нам многократно придётся вычислять, присутствует ли последовательность в транзакциях покупателя. Последовательность может быть достаточно велика, поэтому, для ускорения процесса вычислений, преобразуем последовательности, содержащиеся в транзакциях пользователей в иную форму.
Заменим каждую транзакцию набором последовательностей, которые в ней содержатся. Если в транзакции отсутствуют последовательности, отобранные на предыдущем шаге, то данная транзакция не учитывается и в результирующую таблицу не попадает.
Например, для покупателя с идентификатором 2, транзакция (10, 20) не будет преобразована, поскольку не содержит отобранных последовательностей с нужным значением минимальной поддержки (данный набор встречается только у одного покупателя).
А транзакция (40, 60, 70) будет заменена набором отобранных последовательностей {(40), (70), (40, 70)}
Процесс преобразованная будет иметь следующий вид.

Идентификатор покупателя	Последовательности в покупках	Отобранные последовательности	Преобразованные последовательности
1	<(30)(90)>	<{(30)}{(90)}>	<{1}{5}>
2	<(10, 20)(30)(40, 60, 70)>	<{(30)}{(40)(70)(40, 70)}>	<{1}{2, 3, 4}>
3	<(30, 50, 70)>	<{(30)(70)}>	<{1, 3}>
4	<(30)(40, 70)(90)>	<{(30)}{(40)(70)(40, 70)}{(90)}>	<{1}{2, 3, 4}{5}>
5	<(90)>	<{(90)}>	<{5}>

Фаза генерации последовательностей - из полученных на предыдущих шагах последовательностей строятся более длинные шаблоны последовательностей.
Фаза максимизации - среди имеющихся последовательностей находим такие, которые не входят в более длинные последовательности.

Пусть после фазы трансформации имеется таблица с последовательностями покупок для пяти покупателей.

<{1,5}{2}{3}{4}>

<{1}{3}{4}{3,5}>

<{1}{2}{3}{4}>

<{1}{3}{5}>

<{4}{5}>

Значение минимальной поддержки выберем 40% (последовательность должна наблюдаться как минимум у двоих покупателей из пяти).
После фазы отбора кандидатов мы получили таблицу с одно-элементными последовательностями.

1-Последовательность L_1	**Поддержка**
<1>	4
<2>	2
<3>	4
<4>	4
<5>	4

В фазе генерации последовательностей из исходных одно-элементных последовательностей сгенерируем двух-элементные и посчитаем для них поддержку. Оставим только те, поддержка которых больше минимальной. После этого сгенерируем трёх, четырёх и т.д. элементные последовательности, пока это будет возможно.

2-Последовательность L_2	**Поддержка**
<1 2>	2
<1 3>	4
<1 4>	3
<1 5>	3
<2 3>	2
<2 4>	2
<3 4>	3
<3 5>	2
<4 5>	2

3-Последовательность L_3	Поддержка
<1 2 3>	2
<1 2 4>	2
<1 3 4>	3
<1 3 5>	2
<2 3 4>	2

4-Кандидаты	4-Последовательность L_4	Поддержка
<1 2 3 4>	<1 2 3 4>	2
<1 2 4 3>		
<1 3 4 5>		
<1 3 5 4>		

Последовательность <1 2 4 3>, например, не проходит отбор, поскольку последовательность <2 4 3>, входящая в неё, не присутствует в L_3.
Так как сформировать пяти-элементные последовательности невозможно, работа алгоритма на этом завершается.
Результатом его работы будут три последовательности, удовлетворяющие значению минимальной поддержки и не входящие в более длинные последовательности: <1 2 3 4>, <1 3 5> и <4 5>.

Ограничения AprioriAll

Рассмотренный алгоритм AprioriAll позволяет находить взаимосвязи в последовательностях данных. Это стало возможно после введения на множестве наборов данных отношения порядка (в примере с анализом покупок стало учитываться время транзакции). Тем не менее, AprioriAll не позволяет определить характер взаимосвязи, её силу.

При поиске зависимостей в данных нас могут интересовать только такие, где одни события наступают вскоре после других. Если же этот промежуток времени достаточно велик, то такая зависимость может не представлять значения.

Проиллюстрируем сказанное на примере.

Книжный клуб скорее всего не заинтересует тот факт, что человек, купивший "Основание" Азимова, спустя три года купил "Основатели и Империя". Их могут интересовать покупки, интервал между которыми составляет, например, три месяца.

Каждая совершённая покупка - это элемент последовательности.

Последовательность состоит из одного и более элементов. Во многих случаях не имеет значения, если бы наборы товаров, содержащиеся в элементе последовательности, входили не одну покупку (транзакцию), а составляли бы несколько покупок. При условии, что время транзакций (покупок) укладывалось бы в определённый интервал времени (окно).

Например, если книжный клуб установит значение окна равным одной неделе, то клиент, заказавший "Основание" в понедельник, "Мир-Кольцо" в субботу, и затем "Основатели и Империя" и "Инженеры Мира-Кольцо" (последние две книги в одном заказе) в течении недели, по-прежнему будет поддерживать правило 'Если "Основание" и "Мир-Кольцо", то "Основатели и Империя" и "Инженеры Мира-Кольцо"'.
Ещё одним ограничением алгоритма AprioriAll является отсутствие группировки данных. Алгоритм не учитывает их структуру. В приведённом выше примере можно было бы находить правила, соответствующие не отдельным книгам, а также авторам или литературным жанрам.

Классификация методов

Различают две группы методов:

- статистические методы, основанные на использовании усредненного накопленного опыта, который отражен в ретроспективных данных;

- кибернетические методы, включающие множество разнородных математических подходов.

Статистические методы Data mining

В эти методы представляют собой четыре взаимосвязанных раздела:

- предварительный анализ природы статистических данных (проверка гипотез стационарности, нормальности, независимости, однородности, оценка вида функции распределения, ее параметров и т.п.);
- выявление связей и закономерностей (линейный и нелинейный регрессионный анализ, корреляционный анализ и др.);
- многомерный статистический анализ (линейный и нелинейный дискриминантный анализ, кластерный анализ, компонентный анализ, факторный анализ и др.);
- динамические модели и прогноз на основе временных рядов.

Арсенал статистических методов Data Mining классифицирован на четыре группы методов:

1. Дескриптивный анализ и описание исходных данных.
2. Анализ связей (корреляционный и регрессионный анализ, факторный анализ, дисперсионный анализ).
3. Многомерный статистический анализ (компонентный анализ, дискриминантный анализ, многомерный регрессионный анализ, канонические корреляции и др.).
4. Анализ временных рядов (динамические модели и прогнозирование).

Кибернетические методы Data Mining

Второе направление Data Mining - это множество подходов, объединенных идеей компьютерной математики и использования теории искусственного интеллекта.

К этой группе относятся такие методы:

- искусственные нейронные сети (распознавание, кластеризация, прогноз);
- эволюционное программирование (в т.ч. алгоритмы метода группового учета аргументов);
- генетические алгоритмы (оптимизация);
- ассоциативная память (поиск аналогов, прототипов);
- нечеткая логика;
- деревья решений;
- системы обработки экспертных знаний.

http://www.kdnuggets.com/

Дескриптивные (или описательные) статистики являются базовым и наиболее общим методом анализа данных.

Представьте, что вы проводите опрос с целью составления портрета потребителя товара. Респонденты указывают свой пол, возраст, семейное и профессиональное положение, потребительские предпочтения и т.д., а описательные статистики позволяют получить информацию, на основе которой будет строиться весь портрет. В дополнение к числовым характеристикам создаются разнообразные графики, помогающие визуально представить результаты опроса. Всё это многообразие вторичных данных объединяется понятием «дескриптивный анализ». Полученные в ходе исследования числовые данные наиболее часто представляются в итоговых отчетах в виде частотных таблиц. В таблицах могут быть представлены разные виды частот.

Давайте рассмотрим на примере: *Потенциальный спрос на товар*

Потенциальный спрос на товар

Стоимость товара, руб.	Абсолютная частота, чел.	Относительная частота, %	Кумулятивная частота, %
5000	23	19,2%	19,2%
4500	41	34,2%	53,4%
4399	56	46,6%	100%

Абсолютная частота показывает, сколько раз тот или иной ответ повторяется в выборке. Например, 23 человека купили бы предложенный товар стоимостью 5000 руб., 41 человек – стоимостью 4500 руб. и 56 человек – 4399 руб.

Относительная частота показывает, какую долю данное значение составляет от всего объема выборки (23 человека – 19,2%, 41 – 34,2%, 56 – 46,6%).

Кумулятивная или накопленная частота показывает долю элементов выборки, не превышающих определенное значение. Например, изменение процента респондентов, готовых приобрести тот или иной товар при уменьшении цены на него (19,2% респондентов готовы купить товар за 5000 руб., 53,4% — от 4500 до 5000 руб., и 100% — от 4399 до 5000 руб.).

Наряду с частотами, дескриптивный анализ предполагает расчет различных описательных статистик. Соответствуя своему названию, они предоставляют основную информацию о полученных данных. Уточним, использование конкретной статистики зависит от того, в каких шкалах представлена исходная информация. ***Номинальная шкала*** используется для фиксации объектов, не имеющих ранжированного порядка (пол, место жительства, предпочитаемая марка и т.д.). Для подобного рода массива данных нельзя рассчитать каких-либо значимых статистических показателей, кроме *моды* — наиболее часто встречающегося значения переменной. Несколько лучше в плане анализа ситуация обстоит с ***порядковой шкалой***. Здесь становится возможным, наряду с модой, расчет *медианы* – значения, разбивающего выборку на две равные части. Например, при наличии нескольких ценовых интервалов на товар (500-700 руб. руб., 700-900, 900-1100 руб.) медиана позволяет установить точную стоимость, дороже или дешевле которой потребители готовы приобретать или, наоборот, отказаться от покупки. Наиболее богатыми на все возможные статистики являются ***количественные шкалы***, которые представляют собой ряды числовых значений, имеющих равные интервалы между собой и поддающихся

четыре уровня измерения: номинальный, порядковый, интервальный и отношений

Номинальная шкала

Шкала, содержащая только **категории**; данные в ней не могут упорядочиваться, с ними не могут быть произведены никакие арифметические действия.

Номинальные переменные используются только
для **качественной** классификации. Это означает, что данные переменные могут быть измерены только в терминах принадлежности к некоторым, существенно различным классам; при этом вы не сможете определить количество или упорядочить эти классы. Например, вы сможете сказать, что два индивидуума различимы в терминах переменной А (например, индивидуумы принадлежат к разным национальностям). Данные, измеренными в этой шкале, не могут упорядочиваться, с ними не могут быть произведены никакие арифметические действия.

Номинальная шкала состоит из названий, категорий, имен для классификации и сортировки объектов или наблюдений по некоторому признаку.

Для этой шкалы применимы только операции **равно** (=) и **не равно** (≠).

Часто номинальные переменные называют категориальными.

Примеры:

1) Профессия
2) Город проживания
3) Семейное положение
4) Пол

Порядковая шкала

Шкала, в которой числа присваивают объектам для обозначения **относительной позиции** объектов, но не величины различий между ними.
Шкала измерений дает возможность ранжировать значения переменных.
Измерения же в порядковой шкале содержат информацию только о **порядке следования** величин, но не позволяют сказать **насколько** одна величина больше другой, или насколько она меньше другой.
Порядковые переменные иногда также называют **ординальными**.
Для этой шкалы применимы операции: **равно** (=), **не равно** (≠), **больше** (>), **меньше** (<).
Само расположение шкал в следующем порядке: номинальная, порядковая, интервальная является хорошим примером порядковой шкалы.

Примеры:

1) Место (1, 2, 3…), занятое командой на спортивном соревновании.
2) Номер студента в рейтинге успеваемости (1-й, 23-й, и т.д.), при этом неизвестно, насколько один студент успешней другого, известен лишь его номер в рейтинге.
3) Социоэкономический статус семьи (можно утверждать, что верхний средний уровень выше среднего уровня, однако сказать, что разница между ними составляет, например, 20% мы не сможем).

Интервальная шкала

Шкала, разности, между значениями которой могут быть вычислены, однако их отношения не имеют смысла**.**
Интервальные переменные позволяют не только упорядочивать объекты измерения, но и численно выразить и сравнить **различия** между ними. Например, температура, измеренная в градусах Фаренгейта или Цельсия, образует интервальную шкалу. Вы можете не только сказать, что температура 40 градусов выше, чем температура 30 градусов, но и что увеличение температуры с 20 до 40 градусов вдвое больше увеличения температуры от 30 до 40 градусов.
Эта шкала позволяет находить разницу между двумя величинами, обладает свойствами номинальной и порядковой шкал, а также позволяет определить количественное изменение признака.
Номинальная и порядковая шкалы являются дискретными, а интервальная шкала - непрерывной, она позволяет осуществлять точные измерения признака и производить арифметические операции сложения, вычитания, умножения, деления.
Для этой шкалы применимы операции: **равно** (=), **не равно** (≠), **больше** (>), **меньше** (<),**сложения** (+) и **вычитания** (-).

Пример:
Температура воды в море утром - 19 градусов, вечером - 24, т.е. вечерняя на 5 градусов выше, но нельзя сказать, что она в 1,26 раз выше.

Относительная шкала

Шкала, в которой есть определенная **точка отсчета** и возможны **отношения между значениями** шкалы. Относительные переменные очень похожи на интервальные переменные. В дополнение ко всем свойствам переменных, измеренных в интервальной шкале, их характерной чертой является наличие определенной точки абсолютного нуля, таким образом, для этих переменных являются обоснованными предложения типа: X в два раза больше, чем Y. Типичными примерами шкал отношений являются измерения времени или пространства. Например, температура по Кельвину образует шкалу отношения, и вы можете не только утверждать, что температура 200 градусов выше, чем 100 градусов, но и что она вдвое выше. Интервальные шкалы (например, шкала Цельсия) не обладают данным свойством шкалы отношения. Заметим, что в большинстве статистических процедур не делается различия между свойствами интервальных шкал и шкал отношения.
Для этой шкалы применимы операции: **равно** (=), **не равно** (≠), **больше** (>), **меньше** (<), **сложения** (+)**, вычитания** (-), **умножения** (*) и **деления** (/).
Относительные и интервальные шкалы являются **числовыми**.

Примеры:

1) Вес новорожденных детей 4 кг и 3 кг. Первый ребенок в 1,33 раза тяжелее второго.
2) Цена на картофель в супермаркете в 1,2 раза выше, чем на базаре.

Уровни измерения и соответствующие им статистические методы

Уровни измерений

Описательный метод анализа данных	***Номинальный***	***Порядковый***	***Интервальный***	***Отношений***
Распределение частот	+	+	+	+
Доля	+	+	+	+
Процент	+	+	+	+
Пропорция	+	+	+	+
Мода	+	+	+	+
Медиана		+	+	+
Среднее			+	+

Дескриптивные (*описательные*) методы для всех уровней измерения

Данные на любом из уровней измерения можно описывать в терминах:

1)распределения частот, 2)долей, 3)процентов и 4)пропорций.

Распределение частот Приведем пример простейшего демографического вопроса: Укажите свое нынешнее семейное положение (ПРОЧИТАЙТЕ ВСЕ ПУНКТЫ)

Не женат (не замужем) и никогда не был (а) женат (замужем) _____(1)
Официально женат (замужем) , не живем вместе _____(2)
Не женат (не замужем), разведен (а) _____(3)
Не женат (не замужем), вдовец (вдова) _____(4)
Женат (замужем) _____(5)

Окончательный результат подсчета числа ответов по каждой категорий называется распределением частот.

Распределение частот для данных, собранных с помощью этого демографического вопроса, может выглядеть таким образом:

Категория ответа	Количество ответов
Не женат (не замужем) и никогда не был (а) женат (замужем)	5
Официально женат (замужем) , не живем вместе	10
Не женат (не замужем), разведен (а)	6
Не женат (не замужем), вдовец (вдова)	1
Женат (замужем)	28
ВСЕГО	50

Нынешнее семейное положение	Количество ответов
Состоят в браке	22
Не состоят в браке	28
ВСЕГО	50

Нынешнее семейное положение	Количество ответов
Когда-либо состояли в браке	45
Никогда не состояли в браке	5
ВСЕГО	50

подобная перегруппировка данных дает возможность рассматривать семейное положение совокупности респондентов под разным углом зрения.

Доли, проценты, пропорции

Построив распределение частот, вы должны выбрать один из трех типов анализа, который способствовал бы более глубокому пониманию свойств собранных вами данных. К этим трем типам анализа относятся: доли, проценты и пропорции.

Доли. Доля отражает относительную частоту ответов в категории. Она вычисляется делением числа ответов в конкретной категории на общее число ответов по всем категориям.
Рассмотрим распределение частот ответов на вопрос о семейном положении. От 50 респондентов получено 50 ответов на вопрос. 28 участников ответили, что они в настоящий момент женаты (замужем). Доля женатых (замужних) респондентов в выборке составляет 0,56. Вычисляется она следующим образом: Пропорция женатых (замужних) / Число женатых (замужних) = Общее число участников выборки =28/ 50 = 0,56

Нынешнее семейное положение	Частота	Доля
Не женат (не замужем) и никогда не был (а) женат (замужем)	5	0,1
Официально женат (замужем) , не живем вместе	10	0,2
Не женат (не замужем), разведен (а)	6	0,12
Не женат (не замужем), вдовец (вдова)	1	0,02
Женат (замужем)	28	0,56
ВСЕГО	50	1,00

Вы только что увидели три рекламных ролика. Каждому из роликов было дано название до того, как вы их просмотрели. Ниже ролики перечислены в порядке, в котором вы их увидели. Пожалуйста, дайте оценку каждому из рекламных роликов, указав степень своего доверия к их содержанию. Поставьте «1» напротив названия ролика, который показался вам наиболее правдоподобным, «2» - напротив менее правдоподобного ролика, а «3» поставьте напротив ролика, показавшегося вам наименее правдоподобным. Каждая из оценок от «1» до «3» ставится только один раз. Повторения не допускаются.

«Ученый нового столетия» _____

«Мама нового столетия» _____

«Окружающая среда в новом столетии» _____

Частота рангов каждого рекламного ролика на уровне	*Рекламный ролик «Ученый»*	*Рекламный ролик «Мама»*	*Рекламный ролик «Окружающая среда»*
1	38	10	2
2	8	24	18
3	4	16	30
Сумма частот	**50**	**50**	**50**

Частота рангов каждого рекламного ролика на уровне	*Рекламный ролик «Ученый», %*	*Рекламный ролик «Мама», %*	*Рекламный ролик «Окружающая среда», %*
1	76	20	4
2	16	48	36
3	8	32	60
Итого	**100**	**100**	**100**

Анализ данных по столбцам (сверху вниз) указывает на то, что большая часть участников присвоила:
Рекламному ролику под названием «Ученый» - ранг «1» (76%);
Рекламному ролику под названием «Мама» - ранг «2» (48%),
а рекламному ролику под названием «Окружающая среда» - «3» (60%).

Пропорции.
Третий путь суммирования данных на всех уровнях измерения – использование пропорции. Пропорция одного числа X в отношении другого числа Y определяется как X деленное на Y. Слова по отношению к – важная составляющая этого определения. Число, предваряющее по отношению к (в данном случае число X), ставится в числитель дроби, тогда как число после слов по отношению к ставится в знаменатель дроби. Пропорции, как следует из этой математической формулы, дают возможность отчетливо видеть соотношения между относительным размером двух категорий, использованных в анкетном опросе.

пропорцию не состоящих в браке респондентов по отношению к состоящим в браке можно также выразить как 1:1,27

Анализ данных интервального и относительного уровня измерений

Интервальные и относительные шкалы обладают всеми характерными особенностями, присущими номинальным и порядковым шкалам, а также особыми свойствами, не характерными для этих не столь мощных уровней измерения. Следовательно, все количественные и графические методы, используемые для описания и презентации номинальных и порядковых данных, могут быть применены для описания и представления интервальных и относительных данных. Но сила данных интервального и относительного уровней позволяет осуществить дополнительный анализ, невозможный на номинальном и порядковом уровне. Характер и количество шагов, которые следует предпринять перед применением этих дополнительных методов анализа, зависят от того, являются ли полученные данные дискретными или непрерывными.

Дискретные данные Рассмотрим следующий вопрос для оценки. Пожалуйста, дайте оценку рекламному ролику, который вы только что видели. Для выражения своего согласия или несогласия с утверждением «Этот рекламный ролик рассчитан именно на таких людей, как я» воспользуйтесь приведенной ниже шкалой.

Абсолютно согласен _____(1)
Скорее согласен, чем нет _____ (2)
Не могу сказать определенно _____ (3)
Скорее не согласен _____ (4)
Абсолютно не согласен _____ (5)

Непрерывные данные

- Непрерывные данные предоставляют такую возможность для ответа, при которой значения, по крайней мере, теоретически, могут быть как угодно близко расположены друг к другу на числовой шкале. Например, с помощью вопроса «Сколько вам лет?» собираются непрерывные данные. Респондент может ответить, что ему 40, 40 и 1/2, 41, 42 и 1/3 и т.п. Поскольку вопросы для сбора непрерывных данных не предполагают наличия каких-либо заранее установленных и предварительно закодированных категорий, данные перед вычислением распределения процентов и построением столбиковых или круговых диаграмм следует определенным образом организовать. Организация непрерывных данных называется группировкой (или организацией). Процесс группировки осуществляется в определенной последовательности.
- Данные упорядочиваются.
- Определяются число и ширина интервалов категорий.
- Строится распределение частот.

Несгруппированный ряд ответов на вопрос «Сколько вам лет?»

7	28	39	53
9	28	39	53
9	33	39	54
12	34	41	54
12	34	41	54
13	34	41	54
13	34	41	55
13	34	41	57
13	35	41	58
14	36	41	58
16	36	43	58
16	36	43	63
17	36	43	64
19	37	43	64
19	37	44	64
20	37	44	64
21	37	44	68
21	37	44	69
21	37	44	69
21	37	44	73
21	37	44	73
21	37	47	73
26	39	47	75
27	39	52	75
27	39	53	75

Определение количества и ширины интервалов и категорий. Следующий шаг предполагает определение числа и ширины интервалов категорий. От этого зависит способ группировки данных. По каким критериям группируются данные о возрасте и сколько их – 5 или 25? Твердо установленных правил для проведения границ между категориями не существует. Но при определении ширины интервалов и границ между категориями все же следует иметь ввиду, что:

- группировки должны отражать характер данных. Если размах данных (т.е. разность между наибольшим и наименьшим значениями) большой, тогда и ширина интервалов категорий, скорее всего, будет также большой. Данные, изменяющиеся в более узком диапазоне, лучше обобщать с использованием относительно меньших категорий;
- количество групп не должно быть настолько большим, чтобы скрыть наиболее важные особенности данных, и не столь малым, чтобы лишить систему категорий смысла;
- ширина интервала должна быть целым числом и, по возможности, делиться на удобное число , например на 2, 10, 25, 100 и т.;
- интервалы для всех категорий должны быть, по возможности, одинаковой ширины.

5 категорий	8 категорий	16 категорий
1 - 15	1 - 9	1 - 4
16 - 30	10 - 19	5 - 9
31- 45	20 - 29	10 - 14
46 - 60	30 - 39	15 - 19
70 - 79	40 - 49	20 - 24
	50 - 59	25 - 29
	60 - 69	30 - 34
	70 - 79	35 - 39
		40 - 44
		45 - 49
		50 - 54
		55 - 59
		60 - 64
		65 - 69
		70 - 74
		75 - 79

Вычисление среднего сгруппированных данных

	Первый шаг: вычисление середины интервала	**Второй шаг: умножение середины интервала на частоту**	
Возрастная группа	***Середина интервала***	***Частота***	***Произведение середины интервала на частоту***
0 - 9	4,5	3	13,5
10 - 19	14,5	12	174
20 - 29	24,5	12	294,0
30 - 39	34,5	26	897,0
40 - 49	44,5	20	890,0
50 - 59	54,5	13	708,5
60 - 69	64,5	8	516,0
70 - 79	74,5	6	447,0
Всего		**100**	**3940**
		Третий шаг: сложение произведений	
Четвертый шаг: деление суммы произведений на общее число наблюдений			
= 3940 / 100 = 39,4 = $\overline{X}$			

Среднее является очень мощной статистикой. Оно дает возможность представить одним числом множество ответов на вопрос анкеты. Однако, используя среднее, вы должны быть уверены, что усредненный балл действительно представляет тот ряд ответов, на основе которого он был вычислен.

Приведенная ниже таблица иллюстрирует гипотетический ряд данных о намерении приобрести товар, сложившемся после просмотра одного из рекламных роликов.

Данные о намерении приобрести товар

Я бы купил рекламируемый товар	***Рекламный ролик 1: «Ультра» %***	***Рекламный ролик 2: «Власть» %***	***Рекламный ролик 3: «Дети» %***
Абсолютно согласен (1)	20	50	5
Скорее согласен, чем нет (2)	20	0	15
Не могу сказать определенно (3)	20	0	60
Скорее не согласен (4)	20	0	15
Абсолютно не согласен (5)	20	50	5
Среднее	**3,0**	**3,0**	**3,0**

Значения средних намерения купить, сложившегося после просмотра ка-
ждого рекламного ролика, совпадают, несмотря на то, что лежащие в основе
распределения ответов значительно отличаются друг от друга. Ответы после
просмотра рекламного ролика 1 под названием «Ультра» равномерно распределились по всем пяти категориям, тогда ответы на ролик 2 («Власть») приходятся исключительно на края шкалы. Распределение реакций на рекламный ролик 3 («Дети») напоминают то, что мы зачастую называем колоколообразной *кривой нормального распределения – большинство ответов расположены в центре* распределения, и процент ответов уменьшается к краям шкалы. Изучение этого распределения иллюстрирует важнейший аспект среднего: *среднее становится тем менее репрезентативным по отношению к распределению, на основе которого оно вычисляется, чем больше распределение отличается от*
нормальной кривой.

Несмотря на то, что среднее намерения купить товар равняется 3,0 для всех трех роликов, это значение более репрезентативно для распределения реакций на ролик 3 по сравнению с реакциями на ролики 1 и 2. Нельзя утверждать, что среднее ответов после просмотра рекламного ролика 2 составляет 3,0 или определять его как нейтральное, так как , в сущности, ни один из респондентов не дал ему подобной оценки.

Вычисление дисперсии стандартного отклонения на основе несгруппированных и сгруппированных данных

1.Сгруппированные данные:

Первый шаг: вычислите среднее = [(15 * 1) + (45*3) + (30*4) + (70*5):200] = 695 : 200 = 3,48

Количество респондентов	*Значение ответа*	*Отклонение от среднего*	*Квадраты отклонения*	*Квадраты отклонения * частота*
15	1	-2,48	6,15	92,25
45	2	-1,48	2,19	98,55
40	3	-0,4	0,23	9,20
30	4	+0,52	0,27	8,10
70	5	+1,52	2,31	161,70
Итого = 200				
	Второй шаг: вычислите отклонения от среднего	**Третий шаг: возведите в квадрат отклонения от среднего**	**Четвертый шаг: сложите произведения квадратов отклонений на частоту**	
Пятый шаг: дисперсия = сумма квадратов отклонений : (число респондентов – 1) = 369,8 : 199 = 1,86				
Шестой шаг: стандартное отклонение = $\sqrt{}$ дисперсия = $\sqrt{}$ 1,86 = 1,36				

2.Несгруппированные данные:

Первый шаг: вычислите среднее = (2+1+4+5+5+4+4+5+5+5) : 10 = 40 : 10 = 4,0

Номер респондента	*Значение ответа*	**Второй шаг: вычислите отклонения от среднего** Отклонения от среднего	**Третий шаг: возведите отклонения в квадрат** Квадраты отклонений
1	2	- 2	4
2	1	- 3	9
3	4	0	0
4	5	+ 1	1
5	5	+ 1	1
6	4	0	0
7	4	0	0
8	5	+ 1	1
9	5	+ 1	1
10	5	+ 1	1
Всего = 18			
Четвертый шаг : найдите сумму квадратов отклонений			
Пятый шаг: дисперсия = сумма квадратов отклонений : (число респондентов – 1) = 18:9= 2,0			
Шестой шаг : стандартное отклонение = $\sqrt{}$ дисперсия = $\sqrt{}$ 2,0 = 1,42			

Медиана.
Среднее является часто используемой мерой центральной тенденции ряда данных. Дисперсия и стандартное отклонение указывают на разброс значений вокруг среднего, что позволят сделать вывод о том, насколько хорошо
среднее описывает совокупность данных. Помимо среднего существуют еще две
меры центральной тенденции: медиана и мода. (*Причем следует обратить внимание, что использование среднего, медианы и моды зависит от уровня измерения данных. Среднее вычисляется только для интервальных и относительных данных, медиана – для порядковых, интервальных и относительных данных. Мода используется для свертки данных на всех уровнях измерения).*
*Медианой называется значение, располагающее посередине ранжирован*ного ряда данных. Медиана делит ряд данных пополам таким образом, что 50%
значений меньше медианы.

Что использовать – среднее или медиану?

Определение среднего и медианы ряда значений важно и полезно для более глубокого понимания особенностей данных. В целом, среднее является более предпочтительной мерой в силу своих математических свойств и возможности лучше оценивать среднее генеральной совокупности на основе выборочного среднего. Вместе с тем, существуют *две ситуации, когда следует предпочесть медиану.*

Первая ситуация - когда ряд данных содержит одно или несколько экстремальных значений (так называемых «выбросов» - необычно малых или больших значений). Определять медиану в таких случаях предпочтительнее, поскольку значение среднего чрезвычайно чувствительно к наличию выбросов, тогда как медианы – нет. Если имеются экстремальные значения, среднее можно представить очень искаженную картину.

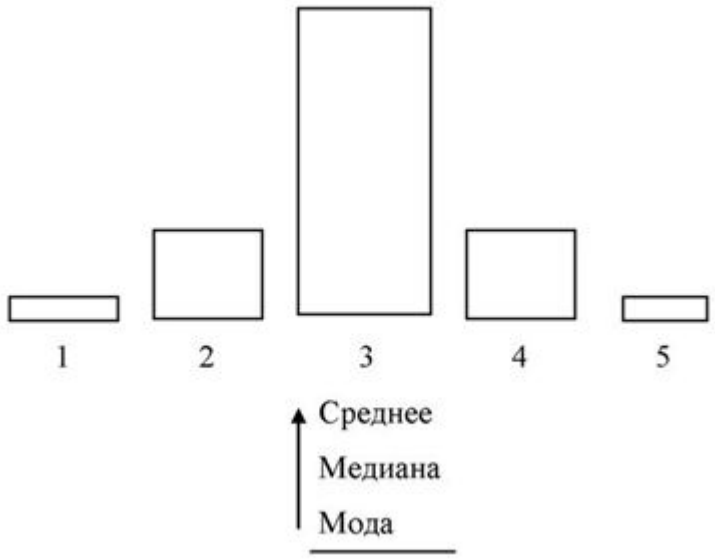
Например, предположим, что вы хотите описать уровень доходов целевой Аудитории нового товара. Вы представляете концепцию нового товара репрезентативной выборке и отмечаете уровни доходов тех, кто сильно или умеренно заинтересован в приобретении товара. Допустим, уровень доходов те

Доход	Частота
10 000	9
12 000	10
17 000	7
20 000	8
747 000	1

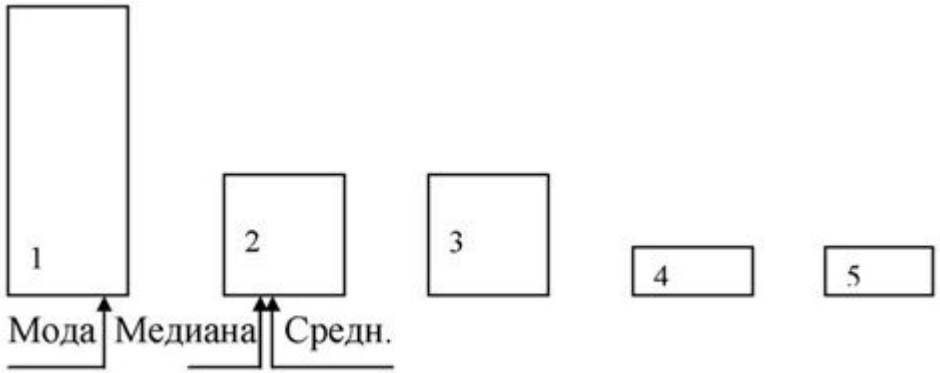
Второй ситуацией, когда следует отдать предпочтение медиане, является наличие открытых категорий в группировке данных. Группировка по возрасту состоит из полностью закрытых групп. Это означает, что каждая возрастная категория имеет верхнюю и нижнюю границу.
Однако для некоторых группировок используются открытые категории. Например, одной из категорий группировки данных о доходах может быть пункт «более 100 тыс. долл.». Среднюю точку этой группы определить *невозможно, так как не установлена верхняя граница. Следовательно, в этой ситуа*ции необходимо использовать *медиану, поскольку без серединной точки вы*числить среднее сгруппированных данных невозможно.

Мода. Еще одной мерой центральной тенденции служит мода. Она определяется как наиболее часто встречающееся значение в ряду данных. Описанные выше шкалы, отражающие намерение купить, имеют *различные моды.* Распределение по рекламному *ролику 1 под названием «Ультра» многомодально, так как существует более двух значений, которые встречаются, которые* встречаются чаще всего. Распределение рекламного ролика под названием «Власть» *бимодально, так как чаще других встречаются два значения. Распр*еделение рекламного ролика под названием «Дети» имеет *одну моду, равную* трем, так как это значение встречается чаще других.

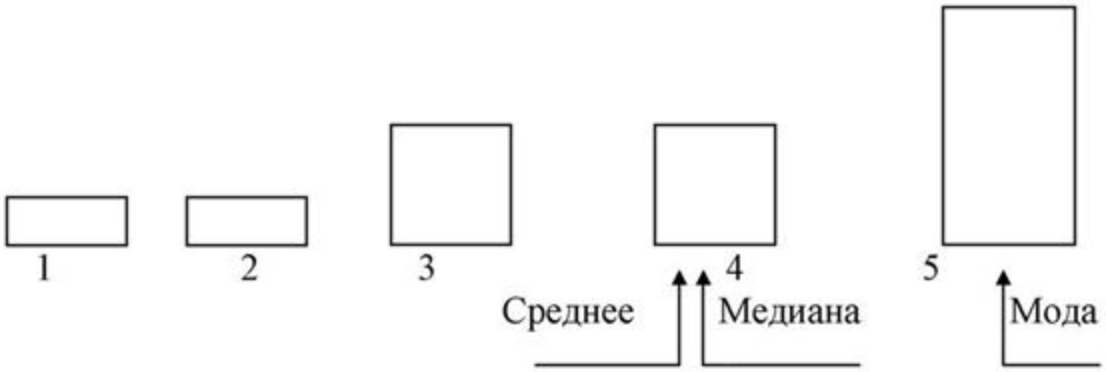
Соотношение среднего, моды и медианы. Среднее, мода и медиана дают различное видение характеристик ряда. Распределение будет симметричным, если среднее, медиана и мода совпадают.



Многие распределения не являются симметричными. Распределение, в котором мода меньше медианы, а медиана в свою очередь, меньше среднего, скошена влево. Это распределение имеет целый ряд значений, с низкой частотой в верхней части.



Распределение, в котором мода больше медианы, а медиана больше среднего, *скошено вправо.*



Упрощенное представление нескольких дескриптивных мер

Номинальный уровень данных: организация представления и вычисление «совокупного» процента.

Вы только что просмотрели рекламный ролик. Поставьте свою отметку напротив утверждения, если вы считаете, что оно отражает именно те чувства, которые вызвал у вас просмотр рекламного ролика. Вы можете отметить сколько угодно утверждений (или вообще не отмечать) в зависимости от чувств, испытанных вами от просмотра рекламного ролика

Было скучно _____
Я кое-что узнал(а), просмотрев рекламный ролик _____
Рекламный ролик рассчитан на таких людей, как я _____
Я видел(а) такие рекламные ролики прежде _____
Лицам, участвующим в рекламном ролике, можно верить _____
Рекламный ролик вызвал у меня замешательство _____
Я скажу своим друзьям, что этот рекламный ролик стоит посмотреть _____
Музыкальное сопровождение прекрасно подобрано _____
Лицу, рекламирующему товар, можно верить _____
Рекламный ролик не интересен _____
Мне не нравятся рекламные ролики такого рода _____
Лицо, рекламирующее товар, вызывает раздражение _____
Хотел(а) бы снова увидеть этот рекламный ролик _____

	Респонденты, выразившие согласие, %
Пример А: порядок пунктов в соответствии с анкетой	
Было скучно	23
Я кое-что узнал(а), просмотрев рекламный ролик	74
Рекламный ролик рассчитан на таких людей, как я	87
Я видел(а) такие ролики прежде	25
Лицам, участвующим в рекламном ролике, можно верить	31
Рекламный ролик вызвал у меня замешательство	26
Я скажу свои друзьям, что этот ролик стоит посмотреть	24
Музыкальное сопровождение прекрасно подобрано	83
Лицу, рекламирующему товар, можно верить	84
Рекламный ролик не интересен	28
Мне не нравятся рекламные ролики такого рода	22
Лицо, рекламирующее товар, вызывает раздражение	26
Хотел(а) бы снова увидеть этот рекламный ролик	78

	Респонденты, выразившие согласие, %
Пример Б: упорядочено по убыванию степени согласия	
Рекламный ролик рассчитан на таких людей, как я	87
Лицу, рекламирующему товар, можно верить	84
Музыкальное сопровождение прекрасно подобрано	83
Хотел(а) бы снова увидеть этот рекламный ролик	78
Я кое-что узнал(а), просмотрев рекламный ролик	74
Лицам, участвующим в рекламном ролике, можно верить	31
Рекламный ролик не интересен	28
Рекламный ролик вызвал у меня замешательство	26
Лицо, рекламирующее товар, вызывает раздражение	26
Я видел(а) такие ролики прежде	25
Я скажу свои друзьям, что этот ролик стоит посмотреть	24
Было скучно	23
Мне не нравятся рекламные ролики такого рода	22

Закономерность ответов на этот вопрос можно сделать более ясной, если
придерживаться следующих действий:

- Во-первых, определите о чем данные будут говорить, т.е. установите,
что вы хотите получить – общую картину положительных или отрицательных откликов, или реакцию на исполнение ролика в сравнении с реакцией на рекламное обращение. (В этом примере мы концентрируем внимание на положительных и отрицательных реакциях).
- Во-вторых, сгруппируйте утверждения в соответствии с целью представления данных. Исходя из поставленной цели, отдельно группируются все положительные утверждения и отдельно – отрицательные.
- В – третьих, дайте название каждой из группировок. В нашем случае одна группировка будет называться «Положительные реакции», а вторая – «Отрицательные реакции».
- В- четвертых, рассчитайте *совокупный процент для каждой группы* суждений. Этот процент характеризует долю респондентов, выбравших, по крайней мере, один из пунктов группировки.

	Респонденты, выразившие согласие, %
Положительные суждения – совокупный процент	**97**
Рекламный ролик рассчитан на таких людей, как я	87
Лицу, рекламирующему товар, можно верить	84
Музыкальное сопровождение прекрасно подобрано	83
Хотел(а) бы снова увидеть этот рекламный ролик	78
Я кое-что узнал(а), просмотрев рекламный ролик	74
Лицам, участвующим в рекламном ролике, можно верить	31
Я скажу свои друзьям, что этот ролик стоит посмотреть	24
Отрицательные суждения – совокупный процент	**31**
Рекламный ролик не интересен	28
Рекламный ролик вызвал у меня замешательство	26
Лицо, рекламирующее товар, вызывает раздражение	26
Я видел(а) такие ролики прежде	25
Было скучно	23
Мне не нравятся рекламные ролики такого рода	22

Когда данные организованы так, как показано в таблице, сразу становятся очевидными следующие выводы:

- Почти всем респондентам что-либо понравилось в рекламном ролике
(учитывая высокий совокупный процент группировки положительных утверждений).
- Большинство потребителей согласились с тем, что рекламный ролик –
именно то, что нужно («рассчитан на таких людей, как я»), а личность, рекламирующая товар, была достаточно убедительной, хотя и вызвала
некоторое раздражение.
- Отрицательные ответы отражают мнение лишь нескольких респондентов
(учитывая низкий совокупный процент группировки негативных утверждений), причем каждому из них не нравится почти все в рекламном ролике.

Интервальные и относительные данные: объединение связанных по смыслу шкал.

Очень часто для оценки индивидуального отношения и поведения используют набор шкальных вопросов. Использование серии шкал обычно обеспечивает многостороннее понимание интересующей области. Например, рекламист, занимающийся репозиционированием товара с целью подчеркнуть его свойства, благотворно влияющие на здоровье человека, сперва может оценить мнение целевой аудитории о рекламировании товаров, благотворно влияющих на здоровье человека, и ее отношение к компаниям, финансирующим такую рекламу. Для этой цели могли быть использованы следующие утверждения:

1. Товар, рекламируемый как «легкий» и «обезжиренный», действительно полезнее для здоровья.
2. Реклама, которая настойчиво подчеркивает свойства товара, благотворно влияющие на здоровье человека, чаще всего простой обман.
3. Корпорации, которые рекламируют свойства товара, благотворно влияющие на здоровье человека, искренне заботятся о потребителе.
4. Реклама, которая настойчиво подчеркивает свойства товара, благотворно влияющие на здоровье человека, эксплуатирует потребности людей.
5. Большинство роликов, которые рекламируют товары, как благотворно влияющие на здоровье человека, малоправдоподобно.
6. Корпорации, которые призывают к потреблению товаров, благотворно влияющих на здоровье человека, стремятся лишь заработать побольше денег.
7. Многие корпорации намеренно преувеличивают свойства своих товаров, представляя их как благотворно влияющие на здоровье чело-

Ответы на утверждения, выражающие отношение

Утверждение*	***Вся выборка***	***Взрослое население 18-25 лет***	***Мужчины 26-54 лет***	***Женщины 26-54 лет***
Товар, рекламируемый как «легкий» и «обезжиренный», действительно полезнее для здоровья.	3,3	3,2	2,9	3,7
Реклама, которая настойчиво подчеркивает свойства товара, благотворно влияющие на здоровье человека, чаще всего простой обман.	4,4	4,6	4,2	4,4
Корпорации, которые рекламируют свойства товара, благотворно влияющие на здоровье человека, искренне заботятся о потребителе	4,1	4,3	3,9	4,1
Реклама, которая настойчиво подчеркивает свойства товара, благотворно влияющие на здоровье человека, играет на слабостях людей	4,5	4,7	4,0	4,8

*Утверждение**	*Вся выборка*	*Взрослое население 18-25 лет*	*Мужчины 26-54 лет*	*Женщины 26-54 лет*
Большинство роликов, которые рекламируют товары, как благотворно влияющие на здоровье человека, малоправдоподобно	4,3	4,7	3,9	4,3
Корпорации, которые призывают к потреблению товаров, благотворно влияющих на здоровье человека, стремятся лишь заработать побольше денег.	4,3	4,4	4,2	4,3
Многие корпорации намеренно преувеличивают свойства своих товаров, представляя их как благотворно влияющие на здоровье человека	4,5	4,7	4,2	4,6

Важные результаты лучше всего представить, сперва организовав утверждения, а затем осуществив дополнительные вычисления. Сначала, как и в случае с вопросами-меню, логически связанные пункты группируются, и группе присваивается название. Далее вычисляется среднее для каждой группы шкал.
Эта обобщающая информация, когда она добавляется в исходную таб-
Лицу «Сгруппированные утверждения, выражающие отношение», делает очевидными и наглядными различия между подгруппами в отношении рекламы и производителей товаров, преподносимых как благотворно влияющие на здоровье человека.

Утверждение	*Вся выборка*	*Взрослое население 18-25 лет*	*Мужчины 26-54 лет*	*Женщины 26-54 лет*
Отношение к рекламированию товара				
Общее отношение	**4,1**	**4,3**	**3,8**	**4,2**
Товар, рекламируемый как «легкий» продукт, ничуть не лучше для здоровья.	3,3	3,2	2,9	3,7
Реклама, которая настойчиво подчеркивает свойства товара, благотворно влияющие на здоровье человека, чаще всего простой обман.	4,4	4,6	4,2	4,4
Реклама, которая настойчиво подчеркивает свойства товара, благотворно влияющие на здоровье человека, играет на слабостях людей	4,5	4,7	4,0	4,8
Большинство роликов, которые рекламируют товары, как благотворно влияющие на здоровье человека, малоправдоподобно	4,3	4,7	3,9	4,3
Отношение к корпорациям				
Общее отношение	**4,3**	**4,5**	**4,1**	**4,3**
Корпорации, которые рекламируют свойства товара, благотворно влияющие на здоровье человека, на самом деле не заботятся о потребителе	4,1	4,3	3,9	4,1
Корпорации, которые призывают к потреблению товаров, благотворно влияющих на здоровье человека, стремятся лишь заработать побольше денег.	4,3	4,4	4,2	4,3
Многие корпорации намеренно преувеличивают свойства своих товаров, представляя их как благотворно влияющие на здоровье человека	4,5	4,7	4,2	4,6

Далее надо иметь ввиду, что усреднение ответов на логически взаимосвязанные шкалы – интуитивно обоснованный метод обобщения информации. Однако для того, чтобы вычисление среднего было осмысленной операцией, вы должны прежде убедиться в том, что шкалы содержательно связаны между собой. Затем следует вычислить коэффициент альфа, который отражает внутреннюю согласованность набора шкал. Среднее арифметической для набора вопросов рекомендуется вычислять только в том случае, если коэффициент альфа для него составляет не менее 0,80.

Вычисление коэффициента альфа

Первый шаг: получите корреляционную матрицу для переменных из набора				
	Переменная 1	Переменная 2	Переменная 3	Переменная 4
Переменная 1	-	-	-	-
Переменная 2	0,876	-	-	-
Переменная 3	0,768	0,769	-	-
Переменная 4	0,963	0,976	0,787	-
Второй шаг: найдите среднее интеркорреляций				
Среднее интеркорреляций = (0,876 + 0,768 + 0,963 + 0,769 + 0,787) : 6 = 5,139 : 6 = 0,857				

Третий шаг: подставьте числа в формулу

$$\textit{Коэффициент альфа} = \frac{N * p}{1 + (p)(N-1)},$$

Где N – это число шкал и p – среднее арифметическое интеркорреляций.

Коэффициент альфа в этом примере составит 0,960, а вычисляется он таким образом:

$$\textit{Коэффициент альфа} = \frac{4 * 0.857}{1 + (0.857)(4-1)}.$$