ФЕДЕРАЛЬНОЕ АГЕНТСТВО ПО ОБРАЗОВАНИЮ

Государственное образовательное учреждение
высшего профессионального образования
Омский государственный университет
им. Ф. М. Достоевского

Исторический факультет
Кафедра социологии

Нагорный Олег Станиславович

Методы text mining в социологии

КУРСОВАЯ РАБОТА
СТУДЕНТА 4 КУРСА

Научный руководитель
кандидат социологических наук
К.В. Павленко

Омск 2014 г.

Содержание

Введение

В обыденном сознании социология у многих ассоциируется со статистикой. Вероятно, это произошло потому, что часто в своих исследованиях социологи, особенно отечественные, применяют достаточно тривиальные методы анализа данных из области классической математической статистики, такие как описательные статистики или анализ таблиц сопряжённости, используя стандартные статистические пакеты, например, SPSS, SAS и другие. Однако наука анализа данных за последние десятилетия ушла далеко вперёд. Лавинообразное увеличение количество информации связанное с развитием Интернета с одной стороны и появление высокопроизводительных компьютеров для её анализа с другой ознаменовали собой начало новой эпохи в дисциплине анализа данных. На основе наработок в области искусственного интеллекта, машинного обучения, статистики и проектировании баз данных в 80-х гг. XX века сформировалась новая междисциплинарная область знания — Data Mining или интеллектуальный анализ данных. Особенность методов, объединяемых данным понятием, заключается в их способности извлекать из «сырых» данных ранее неизвестные нетривиальные знания. Системы Data Mining сейчас находятся на острие исследований и разработок в области анализа, моделирования и практического использования информации и знаний, создавая новую культуры анализа данных.

Сфера применения данных методов практически ничем не ограничена — их можно применять везде, где имеются какие-либо данные [1, стр. 81]. Одной из таких сфер применения является интеллектуальный анализ данных — прежде всего текста — в социальных науках. Группа методов Data Mining, предназначенная для интеллектуального анализа неструктурированного текста объединяется под названием Text Mining. С помощью Text Mining можно получить результаты, недоступные классическим методам анализа данных, например, с высокой точностью спрогнозировать результаты выборов[1] или предсказать популярность фильма до выхода в прокат на основе его обсуждения[2].

Однако по оценке некоторых учёных многие российские социологи не знакомы с данными методами, что нельзя признать нормальным, поскольку «отбрасывает» отечественную социологию на 20-30 лет назад. Отсутствие соответствующей подготовки в области анализа данных приводит к поверхностному анализу эмпирических данных, в то время как важные и полезные неочевидные закономерности в данных «ускользают» от внимания исследователя [2]. Такое игнорирование современных методов анализа данных вполне может стать «фатальной ошибкой»[3] и привести к возникновению «чёрной дыры»[4] в российской социологии. Сказанное позволяет считать, что работа, показывающая перспективы применения методов Data Mining в социологических исследованиях, является **актуальной**.

[1] Прогноз выборов в Венесуэле. URL: http://vox-populi.ru/venezuala.phtml

[2] Predicting the Future With Social Media. URL: http://www.hpl.hp.com/research/scl/papers/socialmedia/socialmedia.pdf

[3] Давыдов А. А. Фатальная ошибка социологии. URL: http://ecsocman.hse.ru/text/28973359/

[4] Орлов А. И. Черная дыра отечественной социологии. URL: http://www.ssa-rss.ru/index.php?page_id=19&id=456

Проблема исследования заключается недостаточности наработок в области применения методов Data Mining в социологии.

Объект исследования — применение методов Data Mining в социологическом исследовании.

Предмет исследования — возможности применения Text Mining для задач классификации и кластеризации неструктурированного текста в социологическом исследовании.

Глава 1

Место text mining в структуре исследовательских методов

1.1 Дуальность статистики

> Если данные говорят с вами, значит вы — байесовец.
>
> Филип А. Шродт [3, стр. 11]

> Bayes' theorem is nominally a mathematical formula. But it is really much more than that. It implies that we must think differently about our ideas.
>
> Нэт Сильвер[1]. The Signal and the Noise.

Дуальность статистики берёт своё начало из философского спора Аристотеля и Платона [4, стр. 7]. Аристотель считал, что реальность может быть познана только эмпирически и что исследователь должен тщательно изучать вещественный мир вокруг себя. Он пришёл к убеждению, что можно разложить сложную систему на элементы, детально описать эти элементы, соединить их вместе и, затем, понять целое. Именно таким механистичным путём долгое время следовала наука. Однако в дальнейшем стало понятно, что не всегда целое можно представить как простую сумму частей, его составляющих. Часто, будучи соединёнными вместе, совокупность этих частей приобретает новое качество.

В отличии от своего ученика, Платон считал что свойством подлинного бытия обладают только идеи, а человек может лишь воспринимать и воплощать в вещах их смутные очертания. Для Платона идея (целое) была большим, чем сумма её материальных проявлений.

Эта дихотомия восприятия реальности проявляется во многих аспектах человеческой мысли, в том числе и в сфере статистического знания, в котором с XVIII в. существует две основных философских позиции относительно того, как применять вероятностные модели. Первая определяет вероятность как нечто, заданное внешним миром. Вторая утверждает,

[1] Американский статистик, давший самые точные прогнозы президентских выборов в США в 2008 и 2012 гг. Входит в 100 самых влиятельных людей в мире по версии журнала Times.

что вероятность существует в головах людей. [5, стр. 18]. В русле первого подхода возникли вначале классическая и затем развивающая её частотная концепции вероятности. Второй подход нашёл выражение в концепции байесовской вероятности.

Сторонники классического подхода исходят из того, что истинные параметры модели не случайны, а аппроксимирующие их оценки случайны, поскольку они являются функциями наблюдений, содержащих случайный элемент. [6, стр. 5-6] Параметры модели считаются не случайными из-за того, что классическое определение вероятности исходит из предположения равновозможности как объективного свойства изучаемых явлений, основанного на их реальной симметрии [7, стр. 24]. На такое представление о вероятности повлияло то, что в начале своего развития теория вероятности применялась прежде всего для анализа азартных игр. Суждение вида «Вероятность выпадения шестёрки при бросании игрального кубика равняется 1/6» основывается на том, что любая из шести граней при подбрасывании на удачу не имеет реальных преимуществ перед другими, и это не подлежит формальному определению. Таким образом, вероятностью случайного события A в её классическом понимании будет называться отношение числа несовместимых (не могущих произойти одновременно) и равновозможных элементарных событий m к числу всех возможных элементарных событий n:

$$P(A) = \frac{m}{n} \tag{1.1}$$

Однако такое определение наталкивается на некоторые непреодолимые препятствия, связанные с тем, что не все явления подчиняются принципу симметрии. Например, из соображений симметрии невозможно определить вероятность наступления дождливой погоды. Для преодоления подобных трудностей был предложен статистический или частотный способ приближённой оценки неизвестной вероятности случайного события, основанный на длительном наблюдении над проявлением или не проявлением события A при большом числе независимых испытаний и поиске устойчивых закономерностей числа проявлений этого события. Если в результате достаточно многочисленных наблюдений замечено, что частота события A колеблется около некоторой постоянной, то мы скажем, что это событие имеет вероятность. Данный тип вероятности был выражен Р. Мизесом в следующей математической формуле:

$$p = \lim_{x \to \infty} \frac{\mu}{n}, \tag{1.2}$$

где μ — количество успешных испытаний, n — количество всех испытаний [7, стр. 46-47]. Вероятность здесь понимается как частота успешных исходов я является чисто объективной мерой, поскольку зависит лишь от точного подсчёта отношения количества успешных и неуспешных событий.

Основываясь на этом подходе, статистика занималась созданием вероятностных моделей, которые включали в себя параметры, которые, как предполагалось, связаны с характеристиками исследуемой выборки. Параметры никогда не могут быть известны с абсолютной точностью до тех пор, пока мы не исследуем всю генеральную совокупность [5, стр. 1]. До тех пор всегда существует вероятность отклонить гипотезу, когда она на самом деле верна, т. е. совершить ошибку первого рода. Для обозначения вероятности такой ошибки частотники используют понятие уровня значимости α. Именно вероятность ошибки первого рода частотники ставят во главу анализа, определяя вероятность вероятность события. После каждого своего утверждения они обычно добавляют «... на доверительном уровне в 95%», подразумевая, что исследователь допускает вероятность ошибки в пяти процентах случаев (при $\alpha = 0,05$) [4, стр. 10-11].

Иногда параметры вообще не возможно интерпретировать применительно к реальной жизни, поскольку модели редко бывают абсолютно верными. Модели, как мы надеемся,

— это некоторые полезные приближения к истине, на основании которых можно делать прогнозы. Тем не менее прежде всего классическое статистическое исследование сосредоточено на оценке параметров, а не на предсказании [5, стр. 1].

Частотный подход доминировал в XX веке, придя на смену другому пониманию вероятности, связанном с именем английского математика Томаса Байеса [8, стр. 2]. Сущность байесовского подхода составляют три элемента: априорная вероятность, исходные статистические данных данные, постаприорная вероятность.

Байесовская статистика начинает построение своей модели при помощи понятия априорной вероятности, с помощью которой описывается текущее состояние наших знаний, относительно параметров распределения [5, стр. 18]. Априорная вероятность, такие образом, — это степень нашей уверенности в том, что исследуемый параметр примет то или иное значение ещё до начала сбора исходных статистических данных. На этом основании байесовское понимание вероятности относят к группе субъективистских трактовок вероятности. Чаще всего предполагается, что для оценки степени уверенности необходимо привлечь экспертов, чьё субъективное свидетельство позволит избежать действительной многократной реализации интересующего нас эксперимента[2] [10, стр. 34].

Следующий элемент — это исходные статистические данные. По мере их поступления статистик пересчитывает распределение вероятностей анализируемого параметра, переходя от априорного распределения к апостериорному, используя для этого формулу Байеса:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}, \tag{1.3}$$

где $P(A)$ — априорная вероятность гипотезы A, $P(A|B)$ — вероятность гипотезы A при наступлении события B (апостериорная вероятность), $P(B|A)$ — вероятность наступления события B при истинности гипотезы A, $P(B)$ — полная вероятность наступления события B. Суть формулы в том, что она позволяет переставить причину и следствие: по известному факту события вычислить вероятность того, что оно было вызвано данной причиной. Эту формулу также называют формулой обратной вероятности. Процесс пересмотра вероятностей, связанных с высказываниями, по мере поступления новой информации составляет существо **обучения на опыте**[3] [6, стр. 21-22] и является одним из возможных способов формализации и операционализации следующего тезиса: «*степень нашей разумной уверенности в некотором утверждении (касающемся, например, неизвестного численного значения интересующего нас параметра) возрастает и корректируется по мере пополнения имеющейся у нас информации относительно исследуемого явления*» [11, стр. 93]. В частотном подходе данный тезис интерпретируется в свойстве состоятельности оценки неизвестного параметра: чем больше объём выборки, на основании которой мы строим свою оценку, тем большей информацией об этом параметре мы распо-

[2]Не следует путать субъективный характер байесовской вероятности в целом с внутренним разделением сторонников данного подхода на объективистов и субъективистов, основанном на различном отношении к роли рациональных ограничений при определении априорной вероятности. В качестве примера различного подхода к определению априорной вероятности рассмотрим ситуацию, где событием является изъятие мячика из урны, наполненной красными и чёрными мячиками — и это всё, что нам известно об урне. Зададим вопрос: какова априорная вероятность (до изъятия мячика), что изъятый мячик будет чёрного цвета? Субъективисты, считающие роль рациональных ограничений относительно небольшой, ответят, что любая вероятность от 0 до 1 может быть рациональной, так как по их мнению наша оценка априорной вероятности зависит большей частью от нерациональных факторов — социализации, свободного выбора и др. Объективисты же будут настаивать, что априорная вероятность в данном случае равняется $1/2$, поскольку именно такая вероятность в соответствии с принципом неопределённости Джейнса инвариантна к к размерам и трансформациям мячиков [9].

[3]Понятие «обучение на опыте» ещё не раз встретится в данной работе, поскольку именно оно составляет суть машинного обучения — подраздела науки искусственного интеллекта, методы которого используются в text-mining.

лагаем и тем ближе к истине наше заключение. Специфика байесовского подхода к интерпретации этого тезиса основана на том, что вероятность, понимаемая как количественное значение степени разумной уверенности в справедливости некоторого утверждения, пересматривается по мере изменения информации, касающейся этого утверждения. Поэтому в данном подходе вероятность всегда есть условная вероятность, при условии нынешнего состояния информации (в русле классического подхода исследователь скорее склонен рассматривать совместную вероятность [4, стр. 5]).

Дискуссии вокруг того, какой же метод предпочтительней, ведутся уже не одно столетие, породив великое множество книг и статей на эту тему [12], [8], но к однозначному выводу прийти не удалось. Острота дискуссии объясняется тем, что спор сторонников байесовского и частотного подхода к статистическому выводу отражает два различных взгляда на способ добычи научного знания. Именно поэтому от ответа на этот, казалось бы, локальный вопрос математической статистики зависит развитие всей науки.

Так или иначе, в 1980-х годах, стало ясно, что частотный подход к статистическому выводу не достаточно хорошо подходит для анализа нелинейных отношений в больших объёмах данных, производимых сложными системами при моделировании процессов реального мира [4, стр. 10]. Для преодоления этих ограничений частотники создали нелинейные версии параметрических методов, такие как множественный нелинейный регрессионный анализ.

В то время как в частотном подходе происходили изменения, немногочисленные сторонники байесовского подхода упрямо продвигали свою точку зрения на модель статистического вывода. Как оказалось, байесовская модель лучше подходит для поиска ответов на некоторые практические вопросы, поскольку полнее учитывает прошлую информацию и располагает к предсказаниям. Например, намного важнее минимизировать вероятность ложноотрицательного диагностирования некоторой опухоли как раковой, чем вероятность её ложноположительного определения (ошибка первого рода).

Продемонстрируем на примере различия в рабоче частотных и байесовских методов проверки гипотез. Предположим, некоторый стрелок утверждает, что точность его стрельбы составляет 75%. Когда стрелка попросили продемонстрировать свои навыки, он попал в мишень только 2 раза из 8. Какова вероятность, что стрелок сказал правду о своих навыках.

Решение задачи в частотном подходе. Гипотеза H_0 — стрелок сказал правду. Испытание — стрельба по мишени. Событие A — попадание в мишень. $P(A)$ постоянная и равна 0,75. Для расчёта вероятности того, что событие A наступило не более 2 раз в 8 независимых испытаниях, применим формулу Бернулли для количества успешных испытаний $k = 0,\ 1,\ 2$ и получим, что $P(A \leq 2) = 0,0042$. Следовательно, при уровне значимости $\alpha = 0{,}05$ следует признать невероятным, что точность стрелка составляет 75%, гипотеза H_0 отвергается.

Отметим некоторые особенности данного решения. Во-первых, для решения задачи мы фактически использовали только умение рассчитывать совместную вероятность, ведь формула Бернулли является сокращённым видом расчёта совместной вероятности успешных комбинаций. Во-вторых, мы решили, что если гипотеза верна, то вероятность отклонить гипотезу, когда она на самом деле верна должна быть не менее 5%, т. е. нам важно, чтобы вероятность ложноположительного ответа была ниже определённой границы. Вероятность ложноотрацательного ответа не рассматривается.

Решение задачи в байесовском подходе. В данном подходе мы не проверяем гипотезу, а рассчитываем условную вероятность события A (точность стрелка составляет 75%) при условии события B (стрелок попал в мишень не более 2 раз из 8). Прежде всего нам нужно оценить априорную вероятность события A. Это можно сделать посмотрев статистику стрельбы остальных стрелков. Предположим, мы выяснили, что 70% стрелков

имеют точность в 75%. Следовательно, $P(A) = 0,7$. $P(B|A)$ мы уже рассчитали в частотном подходе. $P(B)$ легко рассчитывается по формуле полной вероятности. По формуле Байеса $P(A|B) = 0,0301$.

Как видно из этого примера, в байесовском подходе другая логика рассчёт вероятности: на основании данных рассчитывается вероятность того, что H_0 верна, в то время как раньше мы рассчитывали вероятность того, что стрелок поразил мишень не более 2 раз в 8 независимых испытаниях. Данные, полученные с помощью данного метода, данные можно использовать более продуктивно. Предположим, что мы рассчитываем не вероятность того, что стрелок с определённым умениями поразил мишень какое-то количество раз, а вероятность наличия тяжёлого заболевания у человека с каким-то количеством положительных тестов. В случае частотного подхода мы узнаем, какова вероятность того, что больной человек получит н-ое поличество положительных тестов. Байесовский же подход позволяет узнть именно то, что нам надо — вероятность того, что человек, получивший н-ое количество положительных тестов, болен. Другой плюс данных методов — они работают даже если размер выборки равен нулю. В таком случае байесовская вероятность равна априорной.

Проведение тестирования на статистическую значимость оценивает лишь вероятность получения похожего результата с другим набором данных при сохранении тех же самых условий. Однако оно предоставляет ограниченную картину такой вероятности, поскольку в расчет принимается ограниченное количество информации относительно исследуемых данных. И оно само по себе не способно вам сказать, являются ли основные положения исследования верными и будут ли подтверждены полученные результаты в различных условиях [4]. Уровень p говорит только о вероятности получения результата при (обычно) совершенно нереалистичных условиях нулевой гипотезы. А это совсем не то, что мы хотим узнать, — обычно мы хотим знать величину эффекта независимой переменной с учетом имеющихся данных. Это байесовск ий вопрос, а не частотный. Вместо этого значение p часто интерпретируется так, будто бы оно показывало силу ассоциации [3, стр. 11].

С другой стороны у и байесовского метода имеются несколько недостатков. Одним из них является необходимость привлекать для рассчёта априорные данные, которые могут быть недоступны. А если они и доступны, то, как отмечалось выше, часто носят субъективный характер. Другой недостаток — сложность вычислений. В вышеописанном примере для вычисления байесовской вероятности нам необходимо было вычислить частотную вероятность, полную вероятность, и, наконец, собственно байесовскую вероятность. Сложность байесовских вычислений частитчно объясняет тот факт, что байесовские методы вновь обрели популярность с развитием вычислительной техники. Следующий недостаток байесовского метода — неинтуитивность, непонятность его результатов для обыденного сознания. Именно на этой неинтуитивности построен знаменитый парадокс Монти Холла, который легко решает с помощью формулы байеса.

1.2 Data mining как объединение подходов

> В грамм добыча, в годы труды.
> Изводишь единого слова ради
> Тысячи тонн словесной руды.
>
> — В. В. Маяковский

[4] Роль статистической значимости в неудачах науки. URL: http://inosmi.ru/world/20131114/214743342.html

Дальнейшее развитие статистических методов, особенно в их байесовском варианите, привело к возникновению следующего поколения методов статистического анализа, а именно методов машинного обучения. Первоначально эти методы развивались в двух направлениях, первое из которых представлено искусственными нейронными сетями, а второе — деревьями принятия решений [4, стр. 11-12].

Развитие методов машинного обучения в свою очередь привело к созданию статистической теории обучения (Statistical Learning Theory), которая направлена на решения проблемы предсказания на основе имеющихся данных [4, стр. 12-13].

Какое место занимают метды Data Mining в описанной структуре? DM — это междисциплинарная область знания, находящая на пересечении традиционного статистического анализа, искуственного интеллекта, машинного обучения и развития больших баз данных [4, стр. 5]. Можно даже сказать, что DM — это новая философия, новый взгляд на анализ данных.

Хотя как самостоятельная дисциалина DM окончательно оформился в 1990-х гг. [4, стр. 15], о важности ухода от чистой математической статистики в пользу анализа реальных данных говорил ещё Джон Тъюки, который в 1962 году написал статью под названием «Будущее анализа данных» (The future of data analysis), в которой изложил основные идеи новой тенденции. Тьюки говорил о том, что излишняя сосредоточенность на математических теориях в статистики не помогает в решении реальных жизненных проблем. Он был убеждён, что анализ данных — это работа, схожая с работой следователя и что надо дать данным говорить самим за себя. Однако эти идеи тогда не были восприняты приверженцами чистой математической статистики, которые утверждали, что правильная процедура статистического анализа прежде всего предполагает выдвижение научных гипотез, а затем уже из проверку, на основе полученных данных. Попытка анализа данных до выдвижения гипотезы категорически отвергалась, поскольку считалось, что это приведёт к смещению гипотезы в сторону того, что показали данные. Такая позиция привела к тому, что термин «DM» стали использовать в уничижительном значении [13, стр. 788].

Развитие информационных технологий и вычислительной техники с одной стороны привело к появлению огромного количества данных, а с другой — предоставило инструменты для их удобного сбора, хранения и обработки. Эти процессы также изменили течение академических споров, поскольку учёные осознали перспективы новой парадигмы анализа данных. Почему же DM стал популярен в сложившихся условиях?

Суть философии DM частично выражена в названии этой области знания, которое состоит из двух понятий: поиск ценной информации в большой базе данных (data) и добыча горной руды (mining). Именно в просеивании через сито своих инстументов огромного количества «сырых», часто нестуктурированных данных в поисках самородков, т. е. осмысленной, нетривиальной информации — знаний. Более верным названием для этого процесса было бы «knowledge mining from data» (добыча знаний из данных) [14, стр. 5].

Исходное определение термина, которое дал наш бывший соотечественник Григорий Пятнецкий-Шапито, звучит следующим образом: «Data mining — это процесс обнаружения в сырых данных ранее неизвестных нетривиальных практически полезных и доступных интерпритации знаний, необходимых для принятия решений в различных сферах человеческой деятельности» [1, стр. 78].

Литература

1. Дюк В. А., Флегонтов А. В., Фомина И. К. Применение технологий интеллектуального анализа данных в естественнонаучных, технических и гуманитарных областях // Известия Российского государственного педагогического университета им. А.И. Герцена. 2011. № 138. С. 77–84.

2. Давыдов А. А. Knowledge Discovery and Data Mining в системной социологии. 2013. URL: http://www.isras.ru/Davydov_Knowledge.html.

3. Schrodt Philip A. Seven Deadly Sins of Contemporary Quantitative Political Analysis // APSA 2010 Annual Meeting Paper. 2010. URL: http://eventdata.psu.edu/7DS/Schrodt.7Sins.APSA10.pdf.

4. Nisbet Robert, Elder John, Miner Gary. Handbook of statistical analysis and data mining applications. Academic Press, 2009. с. 864.

5. Bayesian Ideas and Data Analysis: An Introduction for Scientists and Statisticians / R. Christensen, W. Johnson, A. Branscum [и др.]. 1 изд. CRC Press, 2010. с. 516.

6. Зельнер А. Байесовские методы в эконометрии. Москва: «Статистика», 1980.

7. Гнеденко Б. В. Курс теории вероятностей. 8 изд. Москва: Едиториал УРСС, 2005.

8. Efron Bradley. Modern Science and the Bayesian-Frequentist Controversy. 2005. URL: http://www-stat.stanford.edu/~ckirby/brad/papers/2005NEWModernScience.pdf.

9. Talbott William. Bayesian Epistemology // The Stanford Encyclopedia of Philosophy / под ред. Edward N. Zalta. 2013.

10. Айвазян С. А., Мхитарян В. С. Прикладная статистика. Основы эконометрики. 2-е, исправленное изд. Москва: Юнити-Дана, 2001. Т. 1. с. 656. URL: http://ecsocman.hse.ru/text/33442857.

11. Айвазян С. А. Байесовский подход в эконометрическом анализе // Прикладная эконометрика. 2008. № 1(9). С. 93–130. URL: http://pe.cemi.rssi.ru/pe_2008_1_93-130.pdf.

12. Jeffreys Harold. Theory of Probability. 3 изд. Oxford: Clarendon Press, 1983.

13. Wilhelm Adalbert. Handbook of Computational Statistics: Concepts and Methods / под ред. J. E. Gentle, W. Härdle, Y. Mori. Springer, 2004. С. 789–803.

14. Han Jiawei, Kamber Micheline. Data Mining: Concepts and Techniques / под ред. Jim Gray. 2 изд. Elsevier, 2006.