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
канд. соц. наук
К.В. Павленко

Омск 2014 г.

Содержание

Введение

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

В. В. Маяковский

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, 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. Произошла целая смена парадигмы [1]

Глава 1

Место text mining в структуре исследовательских методов

1.1 Дуальность статистики

> Если данные говорят с вами,
> значит вы — байесовец.
>
> Филип А. Шродт [2, стр. 11]

Дуальность статистики берёт своё начало из философского спора Аристотеля и Платона [3, 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 отличие 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, что вероятность существует в головах людей. [4, 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. [5, 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 [6, 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 [6, стр. 46-47]. Вероятность в таком понимании — это просто частота. Такой подход также можно назвать дедуктивным, поскольку умозаключение о вероятности события в данном случае делается на основании расчёта доли успешных исходов в длительной серии наблюдений. Вероятность в данном подходе понимается как чисто объективная мера, 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. Более того, 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, не основании которых можно делать прогнозы. Но прежде всего классическое статистическое исследование было сосредоточено на оценке параметров, а не на предсказании [4, стр. 1].

Частотный подход доминировал в XX веке, придя на смену другому пониманию вероятности, связанном с именем английского математика Томаса Байеса [7, 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 [4, 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[1] [9, стр. 34].

[1]Не следует путать субъективный характер байесовской вероятности в целом с внутренним разделением сторонников данного подхода на объективистов и субъективистов, основанном на различном отно-

Следующий элемент — это исходные статистические данные. 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**[2] [5, 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*» [10, 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 располагаем 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, рассматриваемая как степень разумной уверенности в некотором предложении, всегда есть условная вероятность, при условии нынешнего состояния информации (в русле классического подхода исследователь скорее склонен рассматривать совместную вероятность [3, стр. 5]).

Дискуссии вокруг того, какой же метод предпочтительней, ведутся уже не одно столетие, породив великое множество книг и статей на эту тему [11], [7], 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.

1.2 Data mining как продолжение байесовских методов

шении 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 [8].

[2]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.

Литература

1. Давыдов А. А. Knowledge Discovery and Data Mining в системной социологии. 2013. URL: http://www.isras.ru/Davydov$_K$*nowledge.html*.

2. 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.

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

4. Challenges at the Interface of Data Analysis, Computer Science, and Optimization / A. Gaul, A. Geyer-Schulz, L Schmidt-Thieme [и др.]. Springer, 2012. с. 613.

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

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

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

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

9. 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/.

10. Айвазян С. А. Байесовский подход в эконометрическом анализе // Прикладная эконометрика. 2008. № 1(9). С. 93–130. URL: http://pe.cemi.rssi.ru/pe$_2$008$_{19}$3 – 130.*pdf*.

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