Логическая схема байесовского подхода и вероятностного моделирования

Д. Р. Ахмадеев, Е. Е. Хрунова
Финансовый университет при Правительстве Российской Федерации (Финуниверситет), Financial University
akhmadeevdenis@mail.ru, khrunova@ya.ru

***Аннотация*. Байесовский подход, который заключается в том, чтобы найти вероятность события при условии наступления непосредственно с ним связанного, уже больше двухсот лет используется как в разных областях науки, так и в прикладных исследованиях. Благодаря технологическому прогрессу во второй половине двадцатого века Байесовский подход стал еще более актуальным, он реализуется в создании байесовских сетей, представляющий из себя модель (графическую) распределения вероятностей между признаками, которые связаны причинно-следственной зависимостью.**

Ключевые слова: байесовский подход; оценка достоверности; распределение; метод байесовских сетей

I. Введение

Байесовский подход в современном анализе процессов и явлений, привлекая информацию, которая выражается в априорных значениях (заранее известны / основаны на данных статистики/ основаны на мнениях специалистов и экспертов) и совокупности свидетельств, доказывающие или опровергающие нулевые гипотезы, которые относятся к исследуемому объекту, позволяющие через взаимозависимую характеристику элементов составить прогноз вероятности успеха появления ожидаемого события.

Актуальность данной темы заключается в том, что Байесовский подход может помочь в механизмах принятия решений, естественно, при следующих условиях: хорошее качество модели; правильное обоснование и разъяснение полученных результатов.

Сама формула Байеса была опубликована в 1763 году (это спустя 2 года после смерти Томаса Байеса). Как уже выше упоминалось методы, использующие ее, получили действительно широкое распространение только к концу XX века, потому что данные расчеты нуждаются в определенных вычислительных затрат, а их возможно было реализовать только после появления информационных технологий.

Данная работа затрагивает достаточно актуальные вопросы, потому что во время прогресса, развития высоких технологий, увеличения объема информации, которая, в свою очередь, требует структуризацию, то необходимо «очистить» от элементов, отсутствие которых ускорит процесс обработки данных и улучшить само качество их оценки. В экономике, например, данный подход нашел место в рациональном и высоко эффективном осуществлении контроля работы аудиторов, аудиторских организаций саморегулируемыми организациями и Росфиннадзором.

II. Суть и общая логическая схема байесовского подхода

Рассмотрим сначала суть, «философию» данного подхода на примерах.

Допустим, есть модель (закон распределения анализируемой случайной величины, функции регрессии, временного ряда, системы одновременных уравнений и т.п.), где есть s-мерный параметр $\Theta = (\Theta_{1,}\Theta_{2,\ldots,\Theta n})^T$ и нам предстоит построить наилучшую статистическую оценку $\hat{\Theta}$ данного параметра по имеющимся k-мерным наблюдениям $\overline{X}i=(x_i^{(1)},x_i^{(2)},\ldots,x_i^k)^T$, где i принимает значения от 1 до n. T – это верхний индекс, который обозначает операцию транспонирования вектора или же матрицы. Байесовский подход позволяет формализовать и операционализировать тезис, при этом данный тезис принимается что-то вроде абсолютной истины : уровень уверенности (естественно, в ходе разумных и трезвых суждений) в некотором утверждении (например, если речь идет о том, чтобы оценить неизвестное численное значение, параметр которого нас интересует) увеличивается и корректируется по мере возрастания объема информации относительно исследуемого явления. Также допускается существования различных форм интерпретации и подтверждения данного тезиса, при этом даже не имеющие косвенного отношения к байесовскому подходу. Например: свойство состоятельности оценки $\Theta\hat{}$ неизвестного параметра Θ: при увеличении объема выборки, который является основанием для построения оценки $\Theta\hat{}$, у нас увеличивается объем информации и у нас появляется возможность стать ближе к истине, то есть сходимость $\hat{\Theta}_n$ к Θ по вероятности.

Рассмотрим специфику байесовского способа операционализации данного тезиса, который основан на 2 положениях.

1. Степень нашей разумной уверенности» в справедливости некоторого утверждения численно выражается в виде вероятности. Это означает, что вероятность в байесовском подходе выходит за рамки ее интерпретации в терминах условий статистического

ансамбля, но относится к одной из категорий субъективной школы теории вероятностей.)

2. Статистик при принятии решения использует в качестве исходной информации одновременно информацию двух типов: априорную и содержащуюся в исходных статистических данных. При этом априорная информация предоставлена ему в виде некоторого априорного распределения вероятностей анализируемого неизвестного параметра, которое описывает степень его уверенности в том, что этот параметр примет то или иное значение, еще до начала сбора исходных статистических данных. По мере же поступления исходных статистических данных статистик уточняет (пересчитывает) это распределение, переходя от априорного распределения к апостериорному, используя для этого известную формулу Байеса.

$$P\{A_j|B\} = \frac{P\{A_i\}*P\{B|A_i\}}{\sum_{i=1}^{N} P\{B|A_J\}*P\{A_J\}}$$

Данная формула позволяет вычислить условную вероятность события A_i, с учетом того, что произошло некое событие B, именно по безусловной вероятности события A_i и условной вероятностью $P\{B|A_j\}$, где j принимает значения от 1 до N.

Также одним из условием является то, что $A_1,A_2,\ldots,A_N$ – это полная система событий. Второе условие: $P\{B\}>0$, то есть ненулевая вероятность.

Далее изучим реализацию байесовского оценивания неизвестного параметра.

Существуют априорные сведения о некотором параметре Θ. Они, как правило, основаны на опыте функционирования процесса, который анализируется(если есть в условии), и основаны на профессиональном теоретическом соображении о его внутреннем содержании, о его характеристике и его преимуществах и так далее. Суть в том, чтобы эти данные были включены в итоге в виде функции P(Θ), который задает априорное распределение параметра и раскрывает вероятность: параметр будет равен Θ в том случае, если он дискретен; если функция плотности распределения в точке параметра непрерывен изначально.

Исходные статистические данные: $X_1,X_2,\ldots,X_n$ появляются согласно закону распределения вероятностей f(X|Θ), где f(X|Θ) обозначает функцию плотности СВ $\varepsilon=(\varepsilon^{(1)},\varepsilon^{(2)},\ldots,\varepsilon^{(k)})^T$ в т.X, где ε непрерывна; $P\{\varepsilon=X|\Theta\}$ ε дискретная величина, также есть Θ который равен неизвестному параметру.

Также предполагается, что следующие наблюдения X1,X2,…,Xn взаимо-независимы (статистически) при Θ=const. Другими словами, наблюдения образуют случайную выборку из анализируемой генеральной совокупности. Когда мы получаем исходные статистические данные, мы присоединяем эмпирическую информации к имеющейся X априорному сведению, которая представлена как p(Θ).

Функция правдоподобия равна $L(X_1,X_2,\ldots,X_n|\Theta)=f(X_1|\Theta)*f(X_2|\Theta)*\ldots*f(X_n|\Theta)$.

Вычисление апостериорного распределения $\tilde{p}(\Theta|X1, X2,\ldots,Xn)$ решается с помощью байесовского подхода. Здесь A_i событие, которое оценивает параметр Θ, B это событие, которое заключается в том, что все значения n (число наблюдений), произведенных в анализируемой совокупности при $X_1,X_2\ldots.$, которые зафиксированы.

$$\tilde{p}(\Theta|X_1, X_2,\ldots,X_n)=\frac{p(\Theta)L(X1,\ldots,Xn|\Theta)}{\int L(X1,\ldots,Xn|\Theta)*p(\Theta)d\Theta} \quad (1)$$

Построение байесовских точечных и интервальных оценок базируется на применение знаний апостериорного распределения $\tilde{p}(\Theta|X_1, X_2,\ldots,X_n)$, который задается соотношением (1).

В качестве точечных оценок $\hat{\theta}^{(Б)}$ средние/ модальные значения этого распределения:

$$\hat{\theta}^{(Б)}_{(ср)}=E(\Theta|X_1,\ldots,X_n)=\int \Theta\widetilde{p(\Theta|X}\,1,\ldots,\mathrm{Xn})\mathrm{d}\Theta$$

$$\tilde{\theta}^{(Б)}_{мод}=\arg\max\tilde{p}\,(\Theta|X1,\ldots,Xn)$$

Чтобы построить байесовский доверительный интервал для параметра Θ, нужно вычислить по формуле (1) функцию $\tilde{p}(\Theta|X1, X2,\ldots,Xn)$ апостериорного закона распределения Θ, потом с помощью $100*(1+P_0)/2$ определить P_0 и 100*(1-P0)/2-% точки данного закона (они и образуют концы данной интервальной оценки: правый и левый концы).

Важно заметить, что байесовский подход может помочь в точности определения при ограниченном объеме выборки (если сравнивать с классическими подходами-«частотным», например).Тем не менее, при увеличении объема выборки n и классический, и байесовский подход будут давать схожие результаты.

Но в байесовском подходе есть вопросы, которые возникают при практическом применении:

1. Выбор общего вида априорного распределения оцениваемого параметра (параметрическое семейство p(Θ;D). *Как его выбрать?*

2. Выбор численного значения D_0 параметров D, которые определяют некий вид самого априорного распределения уже при выборе, который сделан – общий вид выбора p(Θ;D). *Как подобрать численные значения?*

3. Преодоление вопросов реализации формулы (1), когда вычисляются апостериорное распределение $\tilde{p}(\Theta|X1, X2,\ldots,Xn)$. *Как преодолевать трудности при реализации?*

III. Теоретический аппарат байесовских сетей и его практическое применение

Согласно методологии Байеса, случайность – это мера нашего незнания. Очевидно, что чем больший объем факторов, которые оказывают значительное влияние на

конечный результат мы знаем, тем точнее мы сможем спрогнозировать вероятность его появления.

В трудах французского экономиста и математика Огюстена Курно также прослеживается схожая интерпретация, разделяющая понятия возможности как выражающей нечто объективное и вероятности, которая несет в себе субъективный смысл, не отражая действительно существующего соотношения между вещами, и может оказаться различной для широкого круга лиц в зависимости от объема их знания и незнания.

Получается, что все величины можно измерить, но из-за того, что существует недостаточно полное представление обо всех нюансах параметров, они ведут себя как случайные.

Также если добавить к совокупности данных дополнительные предпочтения или другие параметры, то апостериорное распределение находится по следующей формуле:

$P(A|B)=\frac{P(B|A)P(A)}{P(B)}$, где

- $P(B)=P(B|A)P(A)+P(B|\bar{A})P(\overline{A})$
- $P(B)=\sum_{i=1}^{n} P(B|A)P(A)$,

где P(A) обозначает априорное распределение, то есть вероятность появления конкретного исхода среди остальных возможных; P(B/A) обозначает функцию правдоподобия, то есть совместное распределение выборки из параметрической совокупности, заданной функцией зависимости результата (Y) и факторов, на него влияющих (X).

Вероятность дает возможность предсказывать неизвестные результаты, которые основаны на уже известных параметрах, и сама постановка обратной задачи дается для оценки неизвестных элементов прогноза, который опирается на фактах уже известных результативных показателей.

В первом случае, мы рассматриваем функцию, которая зависит от событий, и во втором, где все зависит именно от параметра, который при фиксированном событии, определяет правдоподобие выбранных величин, которые оказывают влияние на результат;

P(B) обозначает распределение вероятностей (показывает сумму всех возможных исходов данного события);

P(A/B) показывает апостериорное распределение (то есть условное распределение вероятностей какой-либо величины, которое рассматривается как противоположность ее безусловному или априорному распределению).

Эта формула показывает, что оценка вероятностной величины Y может варьироваться зависимо от количества сведений, которые относятся к наблюдаемому и скрытому X, оказывающие влияние на итоговую вероятностную величину.

IV. Пример применения байесовского подхода в современном анализе и моделировании

Проиллюстрируем вышесказанное на экономическом примере, чтобы нагляднее представить применения байесовского подхода в современном анализе.

Пусть A и B равны соответственно H и E,

где H – это экспертная оценка/гипотеза, B – доказательство самой гипотезы.

В этом случае, формула Байеса примет следующий вид:

$P(H_i|E_i\,E_1\ldots E_k)=\frac{P(E1|H_i)P(E_1}{\sum(P(E1\,|\,He)P(E1\,|\,He))\ldots P(Ek\,|\,Hk)P(He)}$

где i принимает значения от 1 до m.

Для начала распределим вероятность событий с помощью правила Байеса, используя базу знаний экспертной системы для того, чтобы вычислить апостериорные вероятности гипотез, учитывая наблюдаемые свидетельства. Это позволит нам узнать сведения об оценочных суждениях, причем для каждого доказательства, который имеется в наличии.

Например, в N имеется три гипотезы:

– H1 (высокая надежность фирмы);

– H2 (средняя надежность фирмы);

– H3 (низкая надежность фирмы).

Априорные вероятности равны $P(H_1)$, $P(H_2)$, $P(H_3)$, соответственно;

Условные независимые события

– E1 (наличие прибыли у фирмы);

– E2 (своевременные расчеты с бюджетом).

Занесем все имеющиеся данные в таблицу № 1 «Распределение вероятностей по заданным условиям».

Дополнительные факты, которые влияют на гипотезу, будут варьировать вероятность, приближая ее значение [0;1] (значение будет зависеть от качества самой взаимосвязи).

Пусть, имеем H1, H2 и H3, и всего лишь одно свидетельство E1, и мы знаем, что его появление достоверное.

ТАБЛИЦА I Распределение вероятностей по заданным условиям

P(i)	1	2	3
P(Hi)	0,3	0,5	0,2
P(E1|Hi)	0,7	0,4	0,1
P(E2|Hi)	0,9	0,6	0

Апостериорное распределение принимает следующий вид:

$P(H_1|E_1)=\frac{P(Ei|Hi)P(Hi)}{\sum_{i=1}^{k} P(Ei|Hi)P(Hi)}$,

где i принимает значения равные 1, 2, 3.

Теперь подставим все значения в формулу:

$P(H_1|E_1)=(0,3*0,7)/(0,3*0,7+0,5*0,4+0,2*0,1)=0,21/0,43=0,49$

$P(H_2|E_2)=(0,5*0,4)/(0,3*0,7+0,5*0,4+0,2*0,1)=0,2/0,43=0,47$

$P(H_3|E_3)=(0,2*0,1)/(0,3*0,7+0,5*0,4+0,2*0,1)=0,02/0,43=0,05$

Итак, вероятность того, что при условии наличия прибыли предприятие будет иметь различные степени надежности:

- высокую 0,49;
- среднюю – 0,47;
- низкую – 0,05.

Можно сделать следующие выводы:

- появляется достоверное событие E1 – доверие к H1 увеличивается;
- к H2 –незначительно сокращается;
- к H3 уменьшается в 4 раза.

Вершины графов – события, которые характеризуются случайной величиной.

Дуги – это вероятностные зависимости, они наглядно демонстрирует возможные варианты распределения показателей. Чтобы правильно дать оценку выбранному параметру, нужно сначала найти следующее отношение: вероятность появления итогового события и остальные подобные значения, появляющиеся на других вершинах, исходящими от других «родителей».

Итак, мы рассмотрели пример, где только одно свидетельство. Теперь добавим к задаче независимое от E1 – событие E2, которое достоверное.

Получаем формулу в следующем виде:

$P(H_1|E_1E_2)=\frac{P(E1|H1)P(E2|H2)P(Hi)}{\sum_{i=1}^{k}P(E1|H1)P(E2\,|\,H2)P(Hi)}$, где i принимает значения равные 1, 2, 3.

В этом случае, апостериорные распределения каждой из гипотез будут равны:

$P(H_1|E_1)=(0,3*0,7*0,9)/(0,3*0,7*0,9+0,5*0,4*0,6+0,2*0,1*0)=0,189/0,309=0,61$

$P(H_2|E_2)=(0,5*0,4*0,6)/(0,3*0,7*0,9+0,5*0,4*0,6+0,2*0,1*0)=0,12/0,309=0,39$

$P(H_3|E_3)=(0,2*0,1*0)/(0,3*0,7*0,9+0,5*0,4*0,6+0,2*0,1*0)=0/0,309=0$

Итак, анализируя полученные результаты, можно сделать следующий вывод.

При условии одновременного появления в вероятностной модели 2-х свидетельств (наличия прибыли; своевременный расчет с бюджетом) – в БЗ остаются только данные гипотезы H1 и H2:

- из них – 61 % принадлежит фирме с высокой степенью надежности,
- 39 % фирме со средней надежностью.

Этот подход можно использовать в разных сферах науки. Также его можно использовать в практике в разных процессах, сопровождающие как сложные модели, так и в бытовой жизни.

V. Заключение

Байесовский подход, который заключается в том, чтобы найти вероятность события при условии наступления непосредственно с ним связанного, уже больше двухсот лет используется как в разных областях науки, так и в прикладных исследованиях. В 21 веке, где происходит технологический прогресс, Байесовский подход стал еще более актуальным, к тому же он реализуется в создании байесовских сетей, который представляет из себя модель (графическую) распределения вероятностей между признаками, которые связаны причинно-следственной зависимостью. Байесовский подход в современном анализе процессов и явлений, привлекая информацию, которая выражается в априорных значениях и совокупности свидетельств, доказывающие или опровергающие нулевые гипотезы, которые относятся к исследуемому объекту, позволяющие через взаимозависимую характеристику элементов составить прогноз вероятности успеха появления ожидаемого события. Это подход до сих пор актуален, потому что число компаний возрастает, где требуется услуга аудиторов, например, а данный подход помогает осуществлять контроль за их деятельностью. Также набирают актуальность консалтинговые компании, где основной задачей является-найти верное и рациональное управленческое решение. Мы живем в век информационных технологий, и окружены бесконечным потоком информации. Мы нуждаемся в «чистке», фильтрации и структуризации информации, чтобы принять правильное решение.

Список литературы

[1] Бондаренко П.С. Теория вероятностей и математическая статистика: учеб. пособие для бакалавров / П.С. Бондаренко, Г.В. Горелова, И.А. Кацко. Краснодар: Кубанский ГАУ 2013. 340 с.

[2] Бережная Е.В., Бережной В.И. Математические методы моделирования экономических систем: Учеб. пособие. М.: Финансы и статистика, 2005. 426 с.

[3] Деловая статистика и вероятностные методы в управлении и бизнесе: учеб. пос. / В.Н. Сулицкий. М.: Изд-во «Дело» АНХ, 2010. 400 с.

[4] Звягин Л. С. Байесовский подход в современном экономическом анализе и имитационном моделировании // Мягкие измерения и вычисления. 2018. № 1. С. 17-26.

[5] Звягин Л.С. Итерационные и неитеративные методы монте-карло как актуальные вычислительные методы байесовского анализа // Международная конференция по мягким вычислениям и измерениям. 2017. Т. 1. С. 39-44.

[6] Звягин Л.С. Системный анализ в социально-экономических и политических системах и применение технологии экспертного прогнозирования // Проблемы конфигурации глобальной экономики XXI века: идея социально-экономического прогресса и возможные интерпретации: Сб. науч. статей. / Под ред. М.Л. Альпидовской, С.А. Толкачева. Краснодар, 2018. С. 184-191.

[7] IBM Watson Health [Электронный ресурс] – Режим доступа:http://www.ibm.com.

[8] Netica Application [Электронный ресурс] – Режим доступа:http://www.norsys.com.

[9] A brief introduction to graphical models and Bayesian Networks[Электронный ресурс] – Режим доступа: https://www.cs.ubc.ca.