Байесовский подход в современном анализе: алгоритмы и синтез

А. В. Савин
Финансовый университет при Правительстве Российской Федерации (Финуниверситет), Financial University
avsavin@yandex.ru

***Аннотация*. Байесовский подход в современном анализе является одним из наиболее распространенных способов формирования аналитической справки о состоянии того или иного объекта на базе апостериорной вероятности, а также построения вывода о возможной реализации той или иной гипотезы касательно состояния данного объекта.**

Ключевые слова: байесовский подход; измерения; вероятность; связи; событие

Ключевыми компонентами байесовского анализа являются функция правдоподобия, которая отражает информацию о параметрах, содержащихся в данных, и априорное распределение, которое количественно определяет информацию о параметрах до начала процесса наблюдения данных. Предварительное (априорное) распределение и вероятность могут быть легко объединены с апостериорным распределением, что представляет собой полное знание параметров после того, как данные были обнаружены. Сравнение априорного и апостериорного значения вероятности могут служить базой для формирования соответствующих аналитических выводов об исследуемом объекте.

Байесовский подход характеризуется существенным отличием от классического подхода к анализу тем, что параметры распределения предполагаются не постоянными, а случайными переменными. Классический подход основан на объективных свидетельствах, в то время как подход Байеса характеризуется субъективной интерпретацией вероятности.

Для байесовского анализа входными данными являются следующие:

- определение переменных системы;
- определение причинно-следственных связей между переменными;
- определение вероятностей (условных и априорных);
- добавление объективных свидетельств к сети;
- обновление доверительных оценок;
- определение апостериорных доверительных точек.

Соответственно, для характеристики выходных данных в теории Байеса мы можем утверждать, что при получении широкого диапазона выходных данных Байесовский метод позволяет использовать полученные данные для построения доверительных интервалов; сети Байеса используются для получения апостериорных распределений.

Байесовский подход представляет собой один из возможных способов процесса формализации и операционализации тезиса, который основан на двух положениях.

1. Степень разумной уверенности в справедливости того или иного утверждения численно выражается в форме вероятности.
2. В ходе принятия решения используется информация двух типов: априорная и содержащаяся в исходных данных. Априорная информация представлена в виде априорного распределения вероятностей анализируемого неизвестного параметра, описывающего степень уверенности принятия параметров того или иного значения до непосредственного сбора статистических данных. По мере получения новых данных статистик уточняет сопутствующие параметры и, соответственно, переходит от априорного предоставления вероятности к апостериорному.

Байесовский метод в современном анализе базируется на формуле Томаса Байеса, английского математика 18 века, который предложил применить сформулированную им теорему для корректировки убеждений с опорой на свежеполученную информацию.

Классическая статистика интерпретирует вероятность как предельную частоту в серии экспериментов повторялась много раз (невозможно сформулировать вероятность неотъемлемо неповторимых событий). В байесовской статистике вероятность представляет собой степень субъективной веры (рационального человека) в истину (истинное / ложное) утверждение. Неопределенность напрямую связана со случайностью, при которой любая неизвестная или неопределенная величина обрабатывается как случайная величина с соответствующим предыдущим распределением.

Базовый принцип байесовский теории предполагает, появление новой информации дает основу для исследователя измерить вероятности, которые обусловлены событиями, связанными между собой. Формула Байеса имеет следующий вид (1).

$$P_a\,H_i = \frac{P(H_i)*P_{Hi}(A)}{\sum_{i=1}^{n} P(H_i)*P_{Hi}(A)} \quad (1)$$

$P(H_i)$ – априорные вероятности (вероятности гипотез до опыта); $P_{Hi}(A)$ – условные вероятности события А при выборе гипотезы i; $P_a(H_i)$ – условная вероятность i-й гипотезы после возникновения события А (апостериорная вероятность).

Следовательно, если рассматривать применение метода Байеса в вероятностном анализе, то мы можем описать процесс как определенную последовательность действий.

1. Априори имеются гипотезы Н о возможных состояниях исследуемого объекта (H_1, H_2….H_n).
2. Данным состояниям присваиваются определенные вероятности на основе статистических данных прошлых лет ($P(H_1)$, $P(H_2)$….$P(H_n)$).
3. Затем проводится эксперимент, в результате которого может наступить или не наступить определенное событие А.
4. Вероятности $P_{Hi}(A)$ определяются опытным путем (возникновение события А при выборе i-й гипотезы).
5. В случае наступления события А проводится замена вероятностей $P(H_i)$ на $P_a(H_i)$ для осуществления переоценки с учетом наступления события А.

Стоит уточнить, что при отсутствии априорных вероятностей применение Байесовского метода является невозможным, так как такого рода формализация становится нецелесообразной для осуществления переоценки вероятности после наступления определенного события.

Далее мы рассмотрим расширенный вариант формулы Байеса, который широко применяется в современном анализе сегодня в различных отраслях. Расширенный вариант формулы Байеса (2) базируется на наличие события Е, которое связано с событиями H_1, H_2….H_n. Вероятности события Е известны в том случае, если какое-либо событие H_1, H_2….H_n наступило: ($P(\mathrm{E}/H_1)$, $P(E/H_2)$….$P(E/H_n)$).

$$P(H_i|E) = \frac{P(\mathrm{E}/H_1)*P(H_i)}{P(\mathrm{E}/H_1)*P(H_1)+P(E/H_2)*P(H_2)+\cdots+P(\mathrm{E}/H_n)*P(H_n)} = \frac{P(EH_i)}{P(E)} \quad (2)$$

H_1, H_2….H_n – гипотезы, событие Е – свидетельство; $P(H_i)$ – вероятности гипотез без учета свидетельства; $P(H_i|E)$ – апостериорные вероятности; $\mathrm{P}(\mathrm{E}H_i)$ – совместная вероятность событий Е и H_i; Р (Е) – полная безусловная вероятность события Е.

Следовательно, мы можем утверждать, что байесовский метод в анализе обладает рядом преимуществ.

1. Для использования данного метода достаточно располагать только априорной информацией.
2. Логически выведенные утверждения и соответствующее им графическое отображение являются доступными для понимания и дальнейшего анализа.
3. Данный метод предполагает использование субъективных вероятностных оценок.

Однако вышеизложенные преимущества метода Байеса в современном анализе также предполагают наличие определенных противоречий, а именно невозможность применения данного метода в анализе некоторых сложных систем, а также необходимость знания множества условных вероятностей, которые получают экспертными методами.

Построение байесовской сети должно начинаться с классификации переменных. Моделирование должно начинаться с идентификации переменных, которые относятся к моделируемой предметной области. Переменные можно разделить на четыре класса в соответствии с их ролью в модели: целевые, свидетельства, факторы, вспомогательные. Целевые переменные отражаются скрытые характеристики и не имеют возможности к прямому измерению (например, неисправность в технической диагностике). Свидетельства представляют собой переменные наблюдения, которые используются для предоставления отображаемой информации о целевых переменных. Факторы являются переменными, которые моделируют источники влияния на целевую переменную, также данные переменные называют контекстными. Факторы делятся на категории (промоутеры, замедлители, требования, исключения) в соответствии с их влиянием на переменную. Промоутеры демонстрируют положительную корреляцию с целевой переменной, замедлители – отрицательную, требования являются обязательным условием к проявлению связанной характеристики, исключения сводят вероятность проявления характеристики к нулю. Вспомогательные переменные используются для построения модели в более доступном виде, к примеру, при наличии у определенного узла родительских узлов, вспомогательные переменные используются для группировки с целью упрощения структуры.

Соответственно, мы можем утверждать, что байесовская сеть – ориентированный граф, который удовлетворяет следующим условиям:

- случайные переменные, которые являются вершинами сети, могут быть как дискретными, так и непрерывными;
- вершины сети соединяются попарно, ориентированными ребрами (к примеру, родительская вершина Х в отношении У, если ребро направлено от вершины Х к вершине У);
- все вершины, которые связаны с родительскими, определяются таблицей условных вероятностей или функцией условных вероятностей;
- для вершин без родителей вероятности состояния являются маргинальными (безусловными).

Соответственно, в силу того, что байесовская сеть представляет собой полную модель для переменных и их отношений, данную модель можно использовать в качестве разрешения вероятностных вопросов. В частности, байесовскую сеть можно использовать с целью получения нового знания о состоянии подмножества переменных в процессе наблюдения за другими

переменными. Подобный процесс вычисления апостериорного распределения переменных по переменным – свидетельствам представляет собой вероятностный вывод. Далее стоит рассмотреть понятие о байесовской классификации, которая выступает альтернативным названием метода байесовских сетей. Байесовская классификация используется для формализации экспертных знаний в экспертных системах.

Широко распространенным сегодня является применение наивно-байесовского подхода, который предполагает изначальную независимость признаков и обладает следующими свойствами:

- все переменные имеют одинаковую значимость;
- все переменные независимы статистически;
- используются все переменные и отображаются все связи между ними.

Данные сети обладают рядом преимуществ.

1. В модели устанавливаются связи между всеми переменными, что предоставляет возможность обрабатывать те случаи, где отсутствуют значения нескольких параметров.
2. Байесовские сети обладают простотой для интерпретации и, соответственно, для построения соответствующих аналитических выводов.
3. Построение байесовской сети способствует совмещению закономерностей, которые были выведены из статистических данных.

Также наивно-байесовский подход обладает рядом недостатков.

1. Перемножение условных вероятностей возможно только тогда, когда переменные действительно являются статистически независимыми. Статистическая независимость в условиях значительного объема входных данных требует применения более комплексных методов и построения сложных многосторонних взаимосвязей.
2. Наивно-байесовский подход предполагает невозможность непосредственной обработки непрерывных переменных, так как требуется их перенесение в интервальную шкалу с целью создания дискретных атрибутов данных переменных.
3. На итоговый результат в наивно-байесовском подходе оказывают влияние только индивидуальные значения входных переменных, так как в данном случае не учитывается комбинирование определенных пар или троек и их взаимное влияние. Учет вышеуказанных комбинаций улучшает качество модели, однако также оказывает влияние на количество и объем необходимой к обработке информации.

Таким образом, в данной части текущей работы нами были описаны теоретические основы байесовского подхода в анализе, а также рассмотрены понятия о байесовских сетях, наивно-байесовской классификации и применение формулы Байеса при решении вероятностных задач. На основе вышесказанного, мы можем прийти к выводу, что байесовский метод позволяет учесть априорные данные о вероятностях с учетом конкретных факторов, включая которые в расчеты можно достичь апостериорных данных о возможности реализации той или иной гипотезы.

Рассматривая байесовские сети, можно отметить, что графическое отображение взаимных взаимосвязей позволяет усилить доступность информации для построения последующих аналитических выводов, так как байесовские сети формируют визуального отображение отношений между переменными, а также отмечают доминантное отношения одной переменной в отношении другой. Одним из плюсов байесовских сетей является также то, что в них возможно включение безусловной переменной.

Прикладное применение байесовского подхода в современном контексте

Байесовские методы используются не только для вероятностного подхода и научно-аналитических исследований. Сегодня определенные отрасли перенимают байесовские принципы с целью создания прикладных инструментов. Например, в современном мире байесовская классификация используется для фильтрации спама. Для эффективной фильтрации спама необходимо соответствие инструмента следующим условиям:

- классифицируемый объект должен обладать определенным количеством признаков (к примеру, слова писем пользователя, которые встречаются с определенной частотой);
- постоянное пополнение набора классификаторов спам-не спам (переобучение).

Данные условия корректно работают в локальных почтовых клиентах, так как поток «не спама» у конечного клиента довольно постоянен, а если изменения присутствуют, то они не носят стремительного характера. Таким образом, подводя итог вышесказанному, мы можем сказать, что выбор байесовских сетей доверия в качестве экспертной системы можно обосновать тем, что логический вывод в байесовских сетях доверия трактуется с точки зрения вычислений, в то время, как системы, которые основаны на теории нечетких множеств не обладают точным обоснованием с математической точки зрения и используют эвристические процедуры.

Использование методологии Байеса в формировании статистических выводов дает возможность совсем по-иному воспринимать и исследовать оцениваемые модели. Он позволяет оперировать не только полученными оценками, а также соответствующими вероятностными распределениями, применять имеющиеся в разных формах априорные знания исследователя относительно оценок параметров модели. Это дает возможность получать большие объемы исходной информации и точнее описывать структуру и другие характеристики исследуемой модели.

Основным преимуществом рассмотренной в практическом примере стратегии выступают простота обработки статистических данных, присутствие

возможности компьютерной реализации принципов (создание прикладных инструментов), возможность накопления новых данных и имплементации элементов пополнения и переобучения и, как следствие, получение актуальных результатов. Таким образом, обобщая, мы можем утверждать, что байесовский подход в переложении на другие отрасли в форме различных прикладных инструментов базируется на постоянном обновлении условных вероятностей и возможности машинного переобучения при появлении новых данных.

Рассмотрим конкретный пример использования байесовского метода в машинном обучении. Формула Байеса может использоваться для выявления случаев и вероятности мошенничества на производственном предприятии при наличии/отсутствии нескольких факторов, косвенно свидетельствующих о наличии мошенничества. Данный алгоритм является самообучаемым, т.е. обладает обратной связью, что свидетельствует о пересчете им полученных коэффициентов при регистрации нового случая мошенничества или его отсутствия. Стоит также отметить, что преобладание классической теории вероятности характерно для многих времен, однако уже в 19 веке значимость учета взаимосвязанности событий в определении конечной вероятности было отмечено многими учеными. Исторически сложилось, что классическая теория вероятностей преобладает в статистическом анализе. На протяжении 19 и 20 веков многие математики, ученые и статистики сопротивлялись использованию байесовской теории. Например, шотландский математик Джордж Кристал настаивал на том, что теорема Байеса и возрождение Лапласа «должны быть прилично скрыты из поля зрения, а не забальзамированы в учебниках и экзаменационных документах». Д-р Эндрю Гельман, профессор статистики в Колумбии, сказал, что «даже если ученые всегда правильно выполняли вычисления, принятие всего с p-значением в 5 процентов означает, что один из двадцати статистически значимых результатов – не что иное, как случайный шум. Доля неправильных результатов, опубликованных в известных журналах, вероятно, даже выше, по мнению Эндрю Гельмана, так как данные часто противоречат друг другу. На основе практического задания, представленного нами в первой части данной работы. Для совершенствования современного применения байесовского метода используются определенные принципы, позволяющие достичь более высокой точности результата и базирующихся на его основе аналитических выводов.

Представление статистической информации в терминах собственных частот, а не вероятностей, повышает производительность в задачах байесовского вывода. Этот положительный эффект естественных частот был продемонстрирован в различных прикладных областях, таких как медицина, право и образование. Однако большинство исследований до сих пор были ограничены ситуациями, когда один дихотомический сигнал используется для определения того, какая из двух гипотез верна. Приложения в реальной жизни часто связаны с ситуациями, когда сигналы (например, медицинские тесты) имеют более одного значения, где рассматриваются более двух гипотез (например, болезней) или где доступно более одной метки. Соответственно, естественные частоты по сравнению с информацией, выраженной в терминах вероятностей, последовательно увеличивают долю байесовских выводов, сделанных статистиками в n условиях. Преподавание собственных частот для простых задач с одной дихотомической репликой и двумя гипотезами приводит к передаче обучения сложным задачам с тремя значениями и большем долей правильных выводов соответственно. Таким образом, естественные частоты облегчают байесовские рассуждения в гораздо более широком классе ситуаций, чем считалось ранее.

К достоинствам байесовского метода в данном контексте можно отнести учет сопутствующих факторов для выявления апостериорной вероятности события с учетом контекста; при этом недостаток проявляет себя в необходимости учета априорных вероятностей, которые при наличии разрозненной сети данных могут обозначаться сугубо экспертными оценками и предоставлять результат, который базируется на данных экспертных оценках, полученных различными методами. Для графического отображения взаимосвязи между переменными существуют байесовские сети, которые позволяют упростить форму представления модели взаимозависимости для построения дальнейших выводов. В частности, байесовскую сеть можно использовать с целью получения нового знания о состоянии подмножества переменных в процессе наблюдения за другими переменными. Подобный процесс вычисления апостериорного распределения переменных по переменным – свидетельствам представляет собой вероятностный вывод.

Также мы можем утверждать, что байесовский подход широко применяется во многих современных отраслях, в частности машинное оборудование, которое базируется на базе байесовской классификации (учет случаев мошенничества на производстве, классификация писем спам-не спам). Более того, вероятностный подход Байеса также используется в технологиях оценки риска и проявляет свое преимущество в возможности создания автоматической системы, которая учитывает новые входные данные для пересчета вероятности (система с обратной связью).

СПИСОК ЛИТЕРАТУРЫ

[1] Айвазян С.А. Байесовский подход в эконометрическом анализе - М.: Синергия, 2008. 332 с.

[2] Звягин Л.С. Процесс обработки информации при реализации концепции "мягких" измерений// Международная конференция по мягким вычислениям и измерениям. 2017. Т. 1. С. 104-109.

[3] Звягин Л.С. Применение системно-аналитических методов в области экспертного прогнозирования// Экономика и управление: проблемы, решения. 2017. Т. 3. № 6. С. 145-148.

[4] Шапиро Л.Д. Экономико-математическое моделирование – Томск: Изд-во Томск. ун-та, 1987. 247 с.

[5] Mayo M., & Mitrovic, A. Optimising ITS behavior with Bayesian networks and decision theory. International Journal of Artificial Intelligence in Education, 2001. 241 p.