Методы анализа случайных данных и их алгоритмизация

Б. Б. Ташполатова
Финансовый университет при Правительстве Российской Федерации (Финуниверситет), Financial University
btash@mail.ru

***Аннотация*. Развитие информационных технологий одновременно сопровождается увеличением количества и разнообразия методов, которые используются для обработки информации. Одним из наиболее эффективных инструментов анализа информации и информационных систем являются вероятностные методы. Они позволяют адекватно описать многие информационные, физические и технологические процессы.**

Ключевые слова: вероятность; подход; информация; стандартизация; оптимизация

В современном мире непрерывно растет поток и объем информации, порожденной наукой, которая превратилась в непосредственную производительную силу.

Характерными чертами информатизации современного общества являются следующие:

1. информация превратилась в важный ресурс производства и привела к снижению потребности в материальных и трудовых ресурсах;
2. информационные технологии вызвали к жизни новые производства;
3. информация превратилась в товар;
4. информация сообщает дополнительную ценность другим ресурсам, например, трудовым.

Причина этого – в хаотическом поведении многих природных объектов и технических систем. Вероятностные методы позволяют с достаточной точностью определить, в каких пределах будет изменяться искомая величина, или с какой вероятностью можно ожидать какого-либо события.

Вероятностно-статистические методы с успехом применяются везде, где существует возможность построить и обосновать вероятностную модель изучаемого процесса или явления.

I. Обработка данных как информационный процесс

Если мы рассмотрим производство информационного продукта, то увидим, как исходный информационный ресурс в соответствии с поставленной задачей в определенной последовательности подвергается различным преобразованиям. Протекающие при этом информационные процессы отражают динамику этих преобразований. Таким образом можно сделать вывод, что информационный процесс – это процесс преобразования информации.

Можно сказать, что обработка информации состоит в получении одних информационных объектов из других информационных объектов путем выполнения некоторых алгоритмов, и является тем самым одной из основных операций, которые осуществляются над информацией, и соответственно выступают главным средством увеличения ее объема и разнообразия.

Можно выделить два вида обработки информации – числовую и нечисловую. При числовой обработке используют такие объекты, как переменные, векторы, матрицы, многомерные массивы, константы и т.д.

В случае нечисловой обработке объектами могут выступать файлы, записи, поля, иерархии, сети, отношения и т.д. В дальнейшем мы будем рассматривать лишь этап обработки данных, а также использование на этом этапе вероятностных методов обработки информации.

II. Статистический анализ данных

Существует два направления статистической обработки данных.

Первое включает методы математической статистики, которые предусматривают возможность вероятностной интерпретации анализируемых данных и получения статистических выводов.

Второе направление объединяет статистические методы, которые исходно не опираются на вероятностную природу обрабатываемых данных. Ко второму подходу обращаются тогда, когда условия сбора исходных данных не укладываются в рамки статистического ансамбля, т.е. в ситуации, когда отсутствует практическая или хотя бы принципиально представимая возможность многократного тождественного воспроизведения базового комплекса условий, при которых осуществлялись измерения анализируемых данных.

Рассмотрим основные этапы обработки данных и кратко охарактеризуем каждый из них. Для этого представим общую логическую схему статистического анализа данных в виде этапов, которые могут быть реализованы в том числе и в режиме итерационного взаимодействия.

На первом этапе происходит предварительный анализ исследуемой системы. На этом этапе определяются: основные цели исследования на неформализованном, содержательном уровне; совокупность единиц (объектов), представляющая предмет статистического исследования; набор параметров-признаков $\left(x^1,\ldots,x^p\right)$ для описания обследуемых объектов; степень формализации соответствующих записей при сборе данных; формализованная постановка задачи.

На втором этапе происходит составление плана сбора исходной информации. При составлении детального плана сбора первичной информации учитывается полная схема анализа. На этом этапе определяется, какой должна быть выборка; объем и продолжительность исследования; схема проведения активного эксперимента (в случае, если он возможен) с привлечением методов планирования эксперимента и регрессионного анализа для определения некоторых входных переменных.

На третьем этапе происходит сбор исходных данных, их подготовка и введение в компьютер для обработки.

Есть два способа представления исходных данных:

- матрица «объект-признак» со значениями k-го признака, который характеризует i-й объект в момент t:

$$x_i^{(k)}(t),\ t=t_1\ldots t_N,\ k=\overline{(1,p)},\ i=\overline{(1,N)};$$

- матрица «объект-объект» $\rho_{ij}(t)$-характеристик попарной близости i-го и j-го объектов или признаков в момент t.

На четвертом этапе производится первичная статистическая обработка данных. При этом решается задача отображения вербальных переменных в номинальную или порядковую (ординальную) шкалу; задача статистического описания исходных совокупностей с определением пределов варьирования переменных; задача анализа резко выделяющихся переменных; задача восстановления пропущенных значений наблюдений; задача проверки статистической независимости последовательности наблюдений, составляющих массив исходных данных; задача унификации типов переменных; экспериментальный анализ закона распределения исследуемой генеральной совокупности и параметризация сведений о природе изучаемых распределений (эту разновидность первичной статистической обработки называют иногда процессом составления сводки и группировки.

На пятом этапе происходит выбор основных методов и алгоритмов статистической обработки данных, составление детального плана вычислительного анализа материала. Пополняется и уточняется тезаурус содержательных понятий. Описывается блок-схема анализа с указанием привлекаемых методов.

На шестом этапе происходит непосредственная реализация плана вычислительного анализа исходных данных.

На седьмом этапе строится формальный отчет о проведенном исследовании. Интерпретируются результаты применения статистических процедур (оценки параметров, проверки гипотез, отображения в пространство меньшей размерности, классификации). При интерпретации могут использоваться методы имитационного моделирования.

III. Основные типы зависимостей между случайными количественными переменными

Под типом зависимости между случайными количественными переменными мы будем понимать не аналитический вид функции $Y_{cp}(X) = f(X,\theta)$, а природу анализируемых переменных (X,y) и, следовательно, интерпретацию функции $f(X,\theta)$.

Наиболее часто рассматривают два типа зависимости – регрессионную и корреляционную.

В первом случае рассматривается регрессионная зависимость случайного результирующего показателя η от неслучайных предсказывающих переменных X.

При этом природа анализируемых связей может носить двойственный характер.

а) замеры показателя η производятся с ошибкой, а замеры неслучайной переменной – X без ошибки.

б) показатель η зависит не только от X, и поэтому для всех X^* значения $\eta(X^*)$ подвержены разбросу.

В данном случае X играет роль параметра, от которого зависит распределение η.

В математическом виде этот случай представляется так

$$\eta(X)=f(X)+\varepsilon(X),$$
$$Y_{\text{ср}}(X)=M\eta(X)=f(X),$$
$$M\varepsilon(X)=0.$$

Мы предполагаем, что природа отклонения $\varepsilon(X)$ и его характеристики распределения не связаны со структурой функции $f(X)$.

Во втором случае рассматривается корреляционно-регрессионная зависимость, которая возникает между случайными векторами η (результирующим показателем) и ξ (объясняющими переменными).

Предполагается, что компоненты векторов в этом случае η и ξ зависят от множества факторов, которые невозможно проконтролировать т.е. эти переменные являются случайными.

Представим η в виде

$$\eta = f(\xi) + \varepsilon$$

где ε – остаточное влияние неучтенных факторов, причем

$$M\varepsilon(k)=0, D\varepsilon(k)=\sigma_k^2<\infty$$
$$cov\,(f(k)(\xi),\varepsilon(k))=0.$$

Для частного случая: m=1; а $f(\xi)$ – линейная функция имеем:

$$\eta = \theta_0 + \sum_{k=1}^{p} \theta_k \cdot \xi^{(k)} + \varepsilon$$

$$Y_{-\mathrm{p}}(x) = \theta_0 + \sum_{k=1}^{p} \theta_k \cdot x^{(k)}$$

Если у нас $\varepsilon = 0$, то случайные величины оказываются связанными чисто функциональной зависимостью $\eta=f(\xi)$, однако ее следует отличать от функциональной зависимости неслучайных переменных.

IV. Методы анализа случайных данных

Рассмотрим основные вероятностные методы анализа данных. Самый распространенный из них – дисперсионный анализ. Возможны несколько реализаций дисперсионного анализа. С учетом числа факторов и имеющихся выборок из генеральной совокупности относительно просто может быть выбран необходимый вариант.

К примеру, однофакторный дисперсионный анализ может быть использован для проверки гипотезы сходства средних значений двух или более выборок, принадлежащих одной и той же генеральной совокупности. Этот метод может быть распространен на тесты двух средних (к которым относится, например, t-критерий).

Двухфакторный дисперсионный анализ с повторениями представляет собой усложненный вариант однофакторного анализа с несколькими выборками для каждой группы данных.

Двухфакторный дисперсионный анализ без повторения – это двухфакторный анализ дисперсии, который включает не более одной выборки на группу. Этот метод анализа может быть использован для проверки гипотезы о равности средних значений двух или нескольких выборок, то есть подтверждения того, что рассматриваемые выборки принадлежат к одной и той же генеральной совокупности.

Корреляционный анализ представляет мощный математический аппарат для количественного определения взаимосвязи двух наборов данных, представленных в безразмерном виде. Под коэффициентом корреляции выборки при этом понимается отношение ковариации двух наборов данных к произведению их стандартных отклонений.

Этот метод дает, например, возможность установить, насколько ассоциированы наборы данных по величине, то есть, насколько большие значения из одного набора связаны с аналогичными по размеру значениями другого набора (случай положительной корреляции), или, наоборот, малые значения одного набора связаны с большими значениями другого (случай отрицательной корреляции), или же данные двух диапазонов никак не связаны между собой (случай нулевой корреляции).

Ковариационный анализ заключается в определении ковариации, которая выступает мерой связи между двумя диапазонами данных. Метод может быть использован для вычисления среднего произведения отклонений точек данных от относительных средних. Он дает возможность установить факт ассоциированности набора данных по величине, то есть, насколько большие значения из одного набора данных связаны с большими значениями другого набора (случай положительной ковариации), или, наоборот, малые значения одного набора связаны с большими значениями другого набора (случай отрицательной ковариации), или же данные двух диапазонов никак не связаны между собой (ковариация близка к нулю).

Двухвыборочный F-тест для дисперсии применяют для решения задачи сравнения дисперсий двух генеральных совокупностей. К примеру, F-тест может быть использован для выявления различий в дисперсиях временных характеристик, которые вычисляются по двум выборкам.

Для решения задач в линейных системах и анализа наборов периодических данных используют Фурье-анализ, основанный на алгоритме метода быстрого преобразования Фурье (БПФ).

Линейный регрессионный анализ заключается в построении методом наименьших квадратов графика, описывающего набор наблюдений. Аппарат регрессионного анализа используется, в частности, для анализа воздействия на отдельную зависимую переменную значений одной или более независимых переменных.

Т-тест может быть использован для проверки средних для различных типов генеральных совокупностей.

Двухвыборочный t-тест Стьюдента служит для проверки гипотезы о равенстве средних для двух выборок.

Двухвыборочный z-тест для средних с известными дисперсиями может быть использован для проверки гипотезы о различии между средними двух генеральных совокупностей.

Алгоритмы, основанные на перечисленных выше методах, входят в большинство стандартных математических пакетов, предназначенных для компьютерной обработки случайных данных.

V. Вероятностные модели информационных открытых систем

Вероятностные методы находят свое применение не только при обработке случайных данных. Так, теория криптографии основана на использовании вероятностных моделей открытых систем.

При решении проблем криптографии используются вероятностные модели информационных открытых систем. При этом система рассматривается в качестве источника случайных последовательностей.

Допустим, что система генерирует в заданном алфавите A_X текст конечной или бесконечной длины. Мы можем в этом случае считать, что источник генерирует конечную или бесконечную последовательность случайных букв $x_0, x_1, x_2, ..., x_i, ...$, которые принимают значения в A_X.

Определим вероятность случайного сообщения $a_0, a_1, a_2, ..., a_{n-1}$ как вероятность последовательности событий:

$$P(a_0, a_1, ..., a_{n-1}) = P(x_0 = a_0, x_1 = a_1, ..., x_{n-1} = a_{n-1}).$$

Множество случайных текстов образует вероятностное пространство, если выполнены условия:

$P(a_0, a_1, a_2, ..., a_{n-1})$ для любого случайного сообщения $a_0, a_1, a_2, ..., a_{n-1}$;

$$\sum_{(a_0, a_1, a_2, ..., a_{n-1})} P(a_0, a_1, a_2, ..., a_{n-1}) = 1;$$

для любого случайного сообщения $a_0, a_1, a_2, ..., a_{n-1}$ и любого $s > n$ справедливо $P(a_0, a_1, a_2, ..., a_{n-1}) = \sum_{(a_s, ..., a_{n-1})} P(a_0, a_1, a_2, ..., a_{s-1})$,

то есть вероятность всех продолжений текста длины n есть сумма вероятностей этого сообщения до длины s.

Текст, порождаемый такой информационной открытой системой, является вероятностным аналогом языка.

Задавая определенное вероятностное распределение на множестве открытых текстов, задается соответствующая модель информационной открытой системы.

Различают стационарные и нестационарные информационные открытые системы. Для стационарных моделей характерно то, что вероятность появления буквы (k-граммы) не зависит от места в открытом тексте.

Рассмотренная модель удобна для практического использования, в то же время некоторые свойства модели противоречат свойствам языка. В частности, согласной этой модели любая k-грамма имеет ненулевую вероятность использования. Вышесказанное не позволяет применять данную модель для дешифрирования широкого класса криптосистем. Вероятностный характер процессов, которые происходят в окружающем нас мире, обусловил интерес исследователей к вероятностным и статистическим методам анализа. Вероятностные методы дают возможность относительно просто построить модель случайного процесса и явления. Математический аппарат этих методов разработан достаточно хорошо, более того, алгоритмы на основе вероятностных методов входят в большинство пакетов математической обработки данных на компьютерах. В настоящей работе было рассмотрено два аспекта использования вероятностных методов в обработке информации. В первом случае рассматривались статистические методов обработки случайных данных. В качестве возможных моделей связи случайных величин была рассмотрена модель регрессии и модель корреляции. Дано описание основных статистических методов описания случайных процессов.

Во втором случае рассматривалось применение вероятностных моделей в теории криптографии.

Показано, что применение вероятностных и статистических методов дает хороший результат как при изучении реальных процессов и систем, так и при проектировании информационных технологий, таких как создание защищенных каналов информации. Обработка экспериментальных данных проводится в целях извлечения из них полезной информации для выработки и принятия управленческих решений. Любая обработка статистических данных – это их преобразование к удобному для использования виду, или перевод ответов исследуемой системы с языка измерений на язык уточняемой модели. Все методы вероятностной обработки информации подразделяются на три большие группы. К первой группе методов относятся наиболее простые и примитивные вычисления средних значений случайной величины за период наблюдения. Ко второй группе можно отнести методы расчетов выборочных характеристик случайной величины. При этом необходимо получение оценка точности и достоверности расчета средних значений. К третьей отнесем методы определения распределения вероятностей случайных величин.

Вероятностные методы обработки информации получили широкое распространение в различных сферах жизни. Особое значение подобные методы приобрели в экономики. Ведь именно в экономической среде поток информации увеличивается с каждым днем, часом, минутой, что неизбежно создает острую необходимость анализа и обработки данных. Однако в мире где столь высока доля неопределённости, человек хоть несколько приблизить себя к изучению определенных фактов, снизить до возможного минимума ошибки при обработке информации. Здесь и проявили особую ценность вероятностные методы обработки информации.

СПИСОК ЛИТЕРАТУРЫ

[1] ГОСТ Р 54500.1-2011/Руководство ИСО/МЭК 98-1:2009. Национальный стандарт Российской Федерации. Неопределенность измерения. Часть 1. Введение в руководства по неопределенности измерения (утв. и введен в действие Приказом Росстандарта от 16.11.2011 N 555-ст)

[2] Походун А.И. Экспериментальные методы исследований. Погрешности и неопределенности измерений: Учебное пособие / А.И. Походун. СПб: СПбГУ ИТМО, 2006. 112 с.

[3] Ефимова Н.Ю. Оценка неопределенности в измерениях: практическое пособие / Н.Ю. Ефимова. Мн.: БелГИМ, 2003. 50 с.

[4] Zvyagin, L.S. The adaptive Bayesian approach to the synthesis of algorithms for joint detection - Estimation. Proceedings of International Conference on Soft Computing and Measurements, SCM 2015. St. Petersburg; Russian Federation; 2015, pp. 18-20. DOI: 10.1109/SCM.2015.7190398.

[5] Zvyagin, L.S. System modeling in marketing research. Proceedings of the 19th International Conference on Soft Computing and Measurements, SCM 2016. St. Petersburg; Russian Federation; 2017, pp. 79-82. DOI: 10.1109/SCM.2016.7519740.