В.В.Нешитой, *доктор технических наук*

ФОРМИРОВАНИЕ НАУЧНОГО МИРОВОЗЗРЕНИЯ СТУДЕНТА ПРИ ИЗУЧЕНИИ МЕТОДОВ СТАТИСТИЧЕСКОГО МОДЕЛИРОВАНИЯ В ИНФОРМАТИКЕ

Глубокое изучение любой дисциплины, допускающей применение количественных методов исследования, должно сопровождаться построением и использованием математических или вероятностно-статистических моделей. Так, в информатике и математической лингвистике широко известны такие модели, как закон Ципфа, применяемый для описания ранговых распределений слов, закон Бредфорда рассеяния информации, закон старения информации и др. К сожалению, каждый из этих законов, как правило, используется сам по себе, без взаимосвязи с другими законами, т.е. без указания его места в ряду других, более общих законов распределения. Кроме того, в литературе имеются такие парадоксальные ситуации, когда статистическое ранговое распределение слов частотного словаря пытаются аппроксимировать (выравнивать) законом Ципфа, а накопленную относительную частоту слов того же частотного словаря – законом Вейбулла, заданным функцией распределения, хотя в обоих случаях необходимо использовать один и тот же закон распределения, а именно тот, который более точно выравнивает статистическое распределение.

Такая путаница возникает в связи с тем, что исследователь берется за решение частной проблемы без предварительного решения общей. Общей же проблемой в данном случае является построение универсальных распределений, объединяющих (включающих как частные случаи) практически все известные распределения, в том числе приведенные выше.

К сожалению, эта проблема в математике не решена. Но, не имея универсальных распределений, мы не можем найти наилучшего выравнивающего распределения. В литературе по теории вероятностей и математической статистике его рекомендуется подбирать исходя из формы статистического распределения, представленного в виде гистограммы. Подбор осуществляется путем выдвижения гипотез о виде теоретической кривой и проверки каждой из них по критериям согласия. Однако такой метод не даст однозначного решения. Кроме того, практика показала, что известные распределения не могут с достаточной точностью описать все многообразие статистических распределений.

О сложности решения задачи по установлению наилучшего теоретического закона распределения свидетельствует тот факт, что порядок обработки статистических данных, выбор математической модели из нескольких предлагаемых частных случаев и в свое время был установлен специальным документом Госстандарта СССР МИ199-79. Этот порядок из-за сложности процедур предлагалось использовать лишь при особо ответственных измерениях, где требовалось определить закон распределения на основании статистических данных, причем необходимость его использования требовалось экономически обосновать.

Выходом из создавшегося положения могло бы быть использование универсальных (многопараметрических, или обобщенных) распределений, которые включали бы как частные случаи все или почти все известные распределения, в том числе семейство кривых К.Пирсона. Тогда не потребовалось бы выдвижения многочисленных гипотез, а наилучшая выравнивающая кривая вычислялась бы за один прием. Но где взять такие универсальные распределения? Ответ может быть один. Построить.

Построение системы непрерывных распределений. Итак, поставим задачу: построить возможно более полную систему

непрерывных распределений, включающую как частные случаи известные классические непрерывные распределения, в том числе семейство кривых К.Пирсона.

Для ее решения необходимо хорошо знать, что такое закон распределения, как он задается аналитически (плотностью вероятностей и функцией распределения), каковы его свойства. Кроме этого, необходимо знать виды случайных величин и их свойства. И, естественно, знать основы высшей математики и уметь использовать эти знания на практике.

Для построения системы непрерывных распределений нам не потребуется доказательства теорем, выдвижения гипотез и т.д. Мы будем выводить все более и более общие законы из самых простых путем введения новых параметров, т.е. будем использовать метод обобщения и другие известные приемы.

Рассмотрим три простейшие непрерывные распределения некоторой случайной величины T: равномерное, треугольное убывающее и треугольное возрастающее. Запишем для каждого из них плотность распределения p(t) и функцию распределения F(t), т.е. интеграл от плотности p(t).

В первом случае имеем

$p(t) = \alpha$, $0<t<1/\alpha$, (1)

$$F(t)=\int_0^t p(t)dt = \int_0^t \alpha dt = \alpha t = 1-(1-\alpha t). \quad (2)$$

Во втором случае

$$p(t)=\alpha(1-\frac{\alpha}{2}t), \quad 0<t<\frac{2}{\alpha}, \quad (3)$$

$$F(t)=\int_0^t \alpha(1-\frac{\alpha}{2}t)dt = 1-(1-\frac{\alpha}{2}t)^2. \quad (4)$$

В третьем случае

$p(t) = 2\alpha t$, $0 < t < \sqrt{1/\alpha}$. (5)

$$F(t) = \int_0^t 2\alpha t dt = \alpha t^2 = 1 - (1 - \alpha t^2). \quad (6)$$

Теперь рассмотрим попарно функции распределения (2), (4) и (2), (6). Формула (4) имеет показатель степени, равный двум. Он появился в результате интегрирования плотности распределения. Обобщим функции распределения (2) и (4) путем введения нового параметра u. Запишем обобщенную (уже с двумя параметрами) функцию распределения в виде:

$$F(t) = 1 - (1 - \alpha u t)^{\frac{1}{u}}. \quad (7)$$

Из формулы (7) при u = 1 имеем формулу (2), а при u = ½ - формулу (4).

Аналогично обобщим формулы (2) и (6) путем введения нового параметра β:

$$F(t) = 1 - (1 - \alpha t^{\beta}). \quad (8)$$

Теперь осталось обобщить формулы (7) и (8):

$$F(t) = 1 - (1 - \alpha u t^{\beta})^{\frac{1}{u}}. \quad (9)$$

Последняя формула включает как частные случаи все предыдущие функции распределения.

Итак, получена трехпараметрическая функция распределения (9). Если ее продифференцировать по t, получим трехпараметрическую плотность распределения:

$$p(t) = \alpha\beta t^{\beta - 1}(1 - \alpha u t^{\beta})^{\frac{1}{u} - 1}. \quad (10)$$

Отнесем полученные трехпараметрические распределения, для которых существуют в явном виде не только плотность, но и функция распределения, к группе А. В других, более сложных случаях интеграл от плотности распределения может не выражаться конечным числом элементарных функций. Поэтому все остальные распределения будем относить к группе Б.

Продолжим процесс обобщения. В формулу (10) параметр β в качестве показателя степени входит дважды. Пусть это будут два разных параметра. Заменим первое вхождение параметра β на параметр γ = kβ. Тогда формулу (10) можно записать в более общем виде:

$$p(t) = Nt^{k\beta-1}(1-\alpha u t^{\beta})^{u^{-1}}, \qquad (11)$$

где N – нормирующий множитель, который находится из условия нормировки:

$$\int_0^{\infty} p(t)dt = 1;$$

α, β, k, u – параметры распределения; t – значение случайной величины T (это может быть ранг слова в частотном словаре, уровень заработной платы и т.д.).

Итак, метод обобщения позволил нам достаточно просто получить обобщенное четырехпараметрическое распределение. Формула (11) включает как частные случаи множество известных распределений, в том числе семейство кривых К.Пирсона. Например, при β = 2, u → 0, kβ = 1 из (11) имеем нормальный закон распределения; при u → 0, k = 1 – закон Вейбулла и т.д. Заметим, что при u → 0 выражение в скобках в формуле (11) имеет вид $e^{\alpha t^{\beta}}$. Это следует из замечательного предела $\lim_{\varepsilon \to 0}(1+\varepsilon)^{1/\varepsilon} = e = 2.71828...$

Плотность (11) предназначена для описания распределений существенно положительных случайных величин, последую-

щие значения которых получаются из предыдущих путем умножения их на некоторую постоянную величину.

Но наряду с такими имеются и другие случайные величины (обозначим их через X), последующие значения которых образуются из предыдущих путем прибавления некоторой постоянной величины (положительной или отрицательной). Найдем обобщенную плотность p(x) для описания таких случайных величин.

В этом случае X = lnT. Перепишем обобщенную плотность (11) в другом виде. Умножим обе части выражения (11) на t и используем запись $e^{\beta \ln t}$ вместо t^{β}, что одно и то же. В результате получим

$$tp(t) = Ne^{k\beta \ln t}(1 - \alpha u e^{\beta \ln t})^{\frac{1}{u}-1}.$$

Вводя обозначения tp(t) = p(x), lnt = x, на основании последней формулы можем записать

$$p(x) = Ne^{k\beta x}(1 - \alpha u e^{\beta x})^{\frac{1}{u}-1}. \quad (12)$$

Формулу (12) можно получить из (11) традиционным методом, т.е. как распределение функции случайного аргумента T. Если X = lnT, то T = e^X. Плотность p(x) получается из известной плотности p(t) по формуле $p(x) = p(t)(dt/dx)$.

Пусть далее T = lnY. Тогда на базе плотности (11) получим новую плотность

$$p(y) = \frac{N}{y}(\ln y)^{k\beta-1}\left(1 - \alpha u(\ln y)^{\beta}\right)^{\frac{1}{u}-1}. \quad (13)$$

И, наконец, при T = lnlnW найдем

$$p(w) = \frac{N}{w \ln w}(\ln \ln w)^{k\beta-1}\left(1 - \alpha u(\ln \ln w)^{\beta}\right)^{\frac{1}{u}-1}. \quad (14)$$

Итак, нами получены четыре системы непрерывных распределений, заданные соответственно плотностями p(x),

p(t), p(y), p(w). Они отличаются между собой началом отсчета значений случайных величин: $X > -\infty$; $T > 0$; $Y > 1$; $W > e$. Эти системы включают как частные случаи множество известных распределений, а также новые, ранее не известные распределения.

Обобщение четырехпараметрических распределений. В начале статьи нами была поставлена задача – построить возможно более полную систему непрерывных распределений. На базе трех простейших распределений: равномерного, треугольного убывающего и треугольного возрастающего – была построена вторая система непрерывных распределений, заданная плотностью p(t). Далее были найдены еще три системы непрерывных распределений как распределения функций случайного аргумента T. Это первая, третья и четвертая системы, заданные соответственно плотностями p(x), p(y), p(w).

Спрашивается, а какие распределения находятся на интервалах между этими четырьмя основными системами?

Чтобы найти эти промежуточные системы распределений, необходимо попарно обобщить плотности p(x), p(t); p(t), p(y); p(y), p(w). Примем за основу обобщения вторую систему непрерывных распределений, заданную плотностью p(t), и рассмотрим три случая.

Случай 1. Введем случайную величину V = f(T), которая задается формулой

$$V = \ln\left(1 + \varepsilon\left(T - \varepsilon\right)\right)^{\frac{1}{\varepsilon}}. \qquad (15)$$

Здесь ε – некоторый параметр, который изменяется на интервале от нуля до единицы, но может принимать и другие значения. Особенностью этой формулы является то, что при $\varepsilon \to 0$ случайная величина V = T, а при ε = 1 V = lnT. Найдем распределение случайной величины V как функции случайного аргумента T. Для этого используем известную формулу $p(v) = p(t)\left(dt/dv\right)$.

Из (15) выразим зависимость T от V

$$T = \frac{e^{\varepsilon V} - 1}{\varepsilon} + \varepsilon. \qquad (16)$$

Далее находим первую производную: $dt/dv = e^{\varepsilon v}$. Тогда на основании плотности p(t) и двух последних равенств имеем

$$p(v) = Ne^{\varepsilon v}\left(\frac{e^{\varepsilon v} - 1}{\varepsilon} + \varepsilon\right)^{k\beta - 1}\left[1 - \alpha u\left(\frac{e^{\varepsilon v} - 1}{\varepsilon} + \varepsilon\right)^{\beta}\right]^{\frac{1}{\alpha u} - 1}. \qquad (17)$$

Здесь $v > \ln(1-\varepsilon^2)^{1/\varepsilon}$. Эта плотность обобщает первую и вторую системы непрерывных распределений. Из (17) при $\varepsilon \to 0$ следует вторая система непрерывных распределений, заданная плотностью (11). Действительно, на основании замечательного предела

$$\lim_{\varepsilon \to 0} \frac{x^{\varepsilon} - 1}{\varepsilon} = \ln x$$

из формулы (16) при $\varepsilon \to 0$ будем иметь $T = \ln e^{V} = V$ и, следовательно, плотность (17) принимает вид (11).

При $\varepsilon = 1$ $T = e^{V}$. В этом случае из формулы (17) следует первая система непрерывных распределений, заданная плотностью (12).

Случай 2. Случайная величина T связана со случайной величиной V зависимостью

$$T = \frac{V^{\varepsilon} - 1}{\varepsilon} + \varepsilon. \qquad (18)$$

Тогда $dt/dv = v^{\varepsilon - 1}$,

$$p(v) = Nv^{\varepsilon - 1}\left(\frac{v^{\varepsilon} - 1}{\varepsilon} + \varepsilon\right)^{k\beta - 1}\left[1 - \alpha u\left(\frac{v^{\varepsilon} - 1}{\varepsilon} + \varepsilon\right)^{\beta}\right]^{\frac{1}{\alpha u} - 1} \qquad (19)$$

Здесь $v > (1-\varepsilon^2)^{1/\varepsilon}$. Эта плотность обобщает вторую и третью системы непрерывных распределений: при ε → 0 из нее следует третья система, заданная плотностью (13), а при ε = 1 – вторая система, заданная плотностью (11).

Случай 3. Случайная величина T связана со случайной величиной V зависимостью

$$T = \frac{(\ln V)^{\varepsilon} - 1}{\varepsilon} + \varepsilon. \qquad (20)$$

Тогда $dt/dv = v^{-1}(\ln v)^{\varepsilon-1}$,

$$p(v) = \frac{N(\ln v)^{\varepsilon-1}}{v}\left(\frac{(\ln v)^{\varepsilon}-1}{\varepsilon}+\varepsilon\right)^{k/\beta-1}\left[1-\alpha t\left(\frac{(\ln v)^{\varepsilon}-1}{\varepsilon}+\varepsilon\right)^{k}\right]^{\frac{1}{m}-1}. \qquad (21)$$

Здесь $v > e^{(1-\varepsilon^2)^{1/\varepsilon}}$.

Эта плотность обобщает третью (при ε = 1) и четвертую (при ε → 0) системы непрерывных распределений.

Итак, мы имеем набор четырехпараметрических систем непрерывных распределений (SNR1, SNR2, SNR3, SNR4) и множество промежуточных систем, т.е. от первой до четвертой системы имеется непрерывный ряд систем, но уже пятипараметрических с дополнительным параметром ε. Эти системы включают как частные случаи великое множество непрерывных распределений.

Напомним, что к этому результату привело обобщение всего лишь трех простейших непрерывных распределений, плотность которых задана уравнением прямой. Полученные четырех-и пятипараметрические распределения убедительно демонстрируют непостижимое волшебство метода обобщения.

Теперь можно быть уверенным, что с вероятностью, близкой к единице, среди этих распределений найдется подходящий частный случай для аппроксимации с достаточной точностью любого статистического распределения, если оно

представляет собой однородную совокупность значений непрерывной случайной величины.

Построенные системы непрерывных распределений могут использоваться не только в информатике и математической лингвистике, но и в других областях знания.

Так, первая система четырехпараметрических распределений хорошо описывает распределение первоисточников по числу цитирований в зависимости от года издания (закон старения информации), а также распределение технологических погрешностей, распределение работников некоторой организации по возрасту.

Вторая система описывает ранговые распределения журналов, упорядоченных по убыванию числа помещенных в них статей по заданному предмету. Из этой же системы выводится математически точная формулировка закона рассеяния информации. Она описывает также распределение слов словаря, фраз и предложений по длине, распределение работающих по уровню заработной платы.

Третья система описывает ранговые распределения полнозначных слов частотного словаря, а также частотных словарей дескрипторов, терминов.

Четвертая система описывает распределение простых чисел.

Закон Ципфа входит как частный случай во вторую и третью системы четырехпараметрических распределений. Закон Вейбулла относится ко второй системе распределений группы А. Из второй системы следуют основные распределения семейства К.Пирсона.

Таким образом, информатика приобрела мощный математический аппарат, а точнее, теорию обобщенных распределений, позволяющую решать множество задач на более высоком уровне. В настоящей статье изложена только часть этой теории, которая используется для моделирования статистических закономерностей текста и информационных потоков. Полная теория включает (кроме рассмотренных непрерывных

распределений) систему дискретных распределений, взаимосвязанную с системой кривых роста новых событий, классификацию распределений, методы и алгоритмы вычисления типа выравнивающей кривой и точечных оценок параметров (универсальный метод моментов и общий устойчивый метод), номограммы для графического определения типа выравнивающей кривой и оценок параметров и серию компьютерных программ под общим названием SNR (системы непрерывных распределений) для работы с указанными системами, а также ряд других программ по всем разделам теории.

Итак, студент вместе с преподавателем на базе трех простейших математических моделей, заданных уравнением прямой, методом обобщения получает новое знание, которое в явном виде не было представлено в исходных моделях. Введение новых параметров в исходные однопараметрические распределения приводит к появлению нового их качества – увеличению аппроксимирующих возможностей за счет возросшего количества форм кривых распределения. При этом студент усваивает не только конкретное новое знание, но и методы его получения.

Решение таких задач способствует формированию научного мировоззрения студентов, учит не только видеть мир, но и глубже понимать его.

1. Нешитой В.В. Исследование статистических закономерностей текста и информационных потоков: дис... докт. техн. навук. – Мн., 1987. – 505 с.

2. Нешитой В.В. Методы статистического анализа на базе обобщенных распределений: учеб.-метод.пособие. – Мн.: Веды, 2001. – 168 с.

3. Нешитой В.В. Статистический анализ и регулирование технологических процессов на базе обобщенных распределений с параметром сдвига: метод. рек. – Мн.: БелГИСС, 2001. – 40 с.