УДК 519.24

В. Ф. Первушин, Н. А. Сергеева, А. В. Стрельников

ПРЕЦИЗИОННЫЙ ГЕНЕРАТОР СЛУЧАЙНЫХ ЧИСЕЛ

Рассматривается алгоритм генерации выборки случайной величины, имеющей заданный закон распределения и численные характеристики (математическое ожидание, дисперсию и т. д.). Показывается высокая точность работы алгоритма на численных примерах моделирования выборки. Общая концепция генерации выборки позволяет выделить данный подход в отдельную категорию алгоритмов генерации случайных чисел, что повлекло объединение таких алгоритмов под названием «прецизионный генератор случайных чисел».

Ключевые слова: случайная величина, закон распределения, числовые характеристики, вероятность события, частота события, объем выборки, гистограмма.

Тему разработок, о которых пойдет речь в данной статье, подсказала задача стохастического моделирования. Если модель построена адекватно физическому процессу, то на ней можно проводить численные эксперименты, связанные с изучением поведения объекта или процесса при изменении входных параметров и влияния внешних воздействии. Также немаловажно более глубоко изучить физические и математические связи внутри самого процесса, найти закономерности и установить меру взаимного влияния параметров и переменных объекта, получить новые качественные знания о процессе, восполнить потерянные или неточные данные. Все эти задачи бывает затруднительно, а порой и невозможно изучать на реальном объекте. Понятно, что некоторые ситуации, если их реализовать на практике, потребуют временных и материальных затрат, изменения хода всего процесса или даже режима работы объекта, чего зачастую допустить нельзя. Именно в таких случаях важность задачи имитационного моделирования становится наиболее значимой. Ведь на модели можно реализовать различные ситуации, в том числе и нетипичные для данного процесса или объекта, и провести наблюдение за его поведением при измененных параметрах входа или зависимостях интересующих исследователей показателей.

Ниже предлагается общая схема процесса, принятая в теории идентификации [1] (рис. 1). Здесь приняты следующие обозначения: $x(t)$ – векторная выходная переменная процесса; $u(t)$ – векторное управляющее воздействие; $\mu(t)$ – векторная входная переменная процесса; $\omega^i(t): i=1,2,...,k$ – переменные процесса, контролируемые по длине объекта; $\xi(t)$ – векторное случайное воздействие; t, заключенное в круглые скобки, – непрерывное время; H со значком вверху – каналы связи, соответствующие различным переменным и включающие в себя средства контроля, приборы для измерения наблюдаемых переменных; значок t внизу переменных x,ω,u,μ означает дискретное время; $h(t)$ – со значком вверху – случайные помехи измерений соответствующих переменных процесса. Контроль переменных x,ω,u,μ осуществляется через интервал времени Δt, т. е. $x_i,\omega_i^1,...,\omega_i^k,u_i,\mu_i$, где $i=\overline{1,s}$ – выборка измерений переменных процесса $\left(x_1,\omega_1^1,...,\omega_1^k,u_1,\mu_1\right)$, $\left(x_2,\omega_2^1,...,\omega_2^k,u_2,\mu_2\right),...,\left(x_s,\omega_s^1,...,\omega_s^k,u_s,\mu_s\right),...$, здесь s – объем выборки.

Материалы, представленные в данной статье, имеют непосредственное отношение к величинам $h(t)$. Речь пойдет о моделировании случайных помех, имеющих определенный закон распределения и числовые характеристики.

Приведем далее некоторые размышления о методе статистического моделирования и его месте в научных исследованиях. Метод статистического моделирования широко используется в разнообразных задачах кибернетики для исследования алгоритмов идентификации, распознавания образов, управления и т. д. В последние годы появилось много эвристических алгоритмов, прежде всего в связи с тем, что новые формулировки задач не поддаются строгой математической постановке. А это, как правило, означает отсутствие процедуры аналитического синтеза тех или иных алгоритмов, доказательство соответствующих теорем сходимости, наличие которых ранее считалось мерой истинности, правильности и основанием того, что дальнейшие действия являются правомерными. В этой связи метод статистического моделирования следует считать доказательным этапом, а не иллюстрацией работы тех или иных алгоритмов. Последнее существенно повышает требования к проведению подобных исследований. Определяющей здесь является возможность повторения определенного цикла численных экспериментов другими исследователями.

Во многих случаях на этом пути безусловно важной является необходимость работы со случайными помехами, распределенными по конкретному закону, а не просто с датчиками случайных чисел, о которых говорят, что они распределены по равномерному, нормальному либо какому-то другому закону. Как оказалось на практике, существующие генераторы случайных чисел весьма условно можно назвать соответствующими заявленным законам распределения. Особенно заметно отклонение на выборках небольших объемов. Все классические алгоритмы генерации [2] верны прежде всего на бесконечно большом объеме выборки, ведь все теоремы сходимости содержат формулировку «Пусть объем выборки стремится в бесконечность» или тому подобное допущение. Однако если объем выборки и достаточен для того, чтобы

служить основой для построения модели, то он все равно далек от того, чтобы применить к нему термин «бесконечный». Это говорит о том, что следует задуматься о создании алгоритмов генерации конечной выборки случайной величины и способ создания такой величины будет существенно отличаться от классического подхода. Мы ни в коем случае не спорим с классикой, однако нам хотелось бы больше внимания уделить задачам практики и увязать классический подход с прикладными задачами.

При исследовании многих стохастических процессов предполагается, что плотность вероятности случайных факторов (помех) подчиняется какому-либо закону. В частности, что ошибка измерения электрических и неэлектрических величин электрическими способами подчиняется нормальному закону, в задачах теории надежности используется распределение Вейбулла (интенсивность отказа), логнормальное распределение характерно для многих физических и социально-экономических ситуаций (размер и вес частиц, образующихся при дроблении; заработная плата работника; размеры космических образований; долговечность изделия, работающего в режиме износа и старения, и др.) [3].

Для того чтобы сгенерировать множество случайных точек, свойства которых подчиняются одному из известных законов распределения, необходимо задать закон распределения и его параметры, количественно характеризующие случайную величину, такие как математическое ожидание, среднеквадратическое отклонение, мода и др., и определить количество случайных точек.

Рассмотрим генерацию множества точек для одномерной случайной величины (рис. 2).

Пусть необходимо сгенерировать статистически независимую выборку случайной величины X с плотностью вероятности распределения $f(x, m_x, \sigma_x, ...)$. Обозначим через n количество генерируемых точек, т. е. объем будущей выборки. Далее зададим интервал $[a, b]$, на котором будет генерироваться выборка.

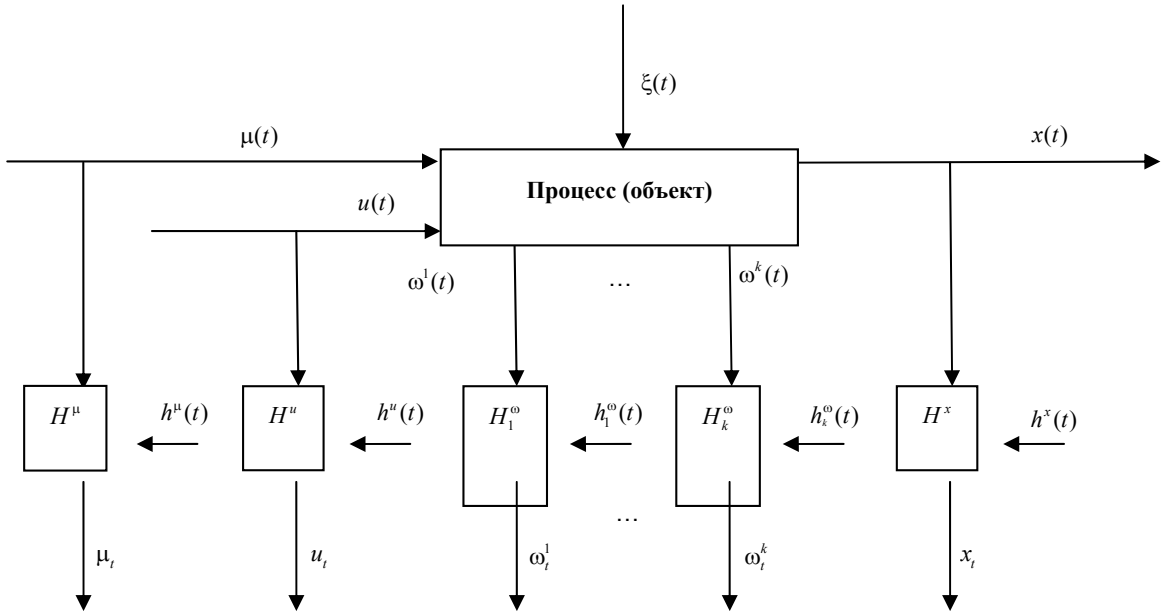

Рис. 1. Схема процесса

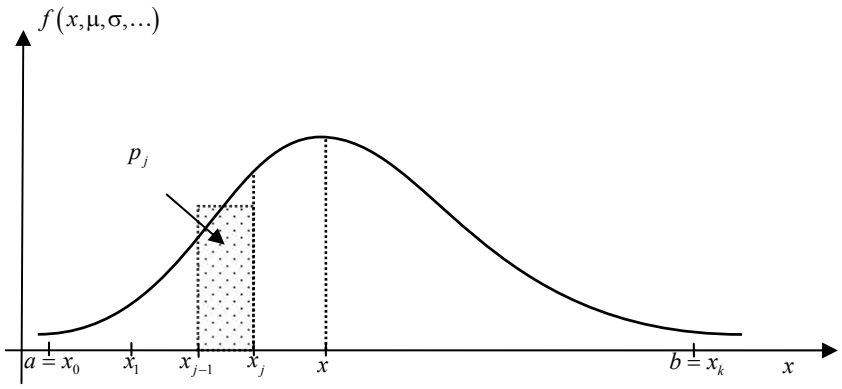

Рис. 2. Графическая интерпретация алгоритма

Выборка генерируемых точек должна наиболее полно охватывать всю область возможных значений, но по понятным причинам следует отрезать «хвосты» распределения, где вероятность выпадения значений X становится малой. В этом случае нужно определить этот порядок малости и, например, принять к получению выборки тот интервал, внутри которого $f(x, m_x, \sigma_x, ...)$ не меньше некоторого значения. Определим это значение пропорционально $f_{\max}(x, m_x, \sigma_x, ...)$:

$$f(x = a, m_x, \sigma_x, ...) = f(x = b, m_x, \sigma_x, ...) = \\ = 0,01 \cdot f_{\max}(x, m_x, \sigma_x, ...). \quad (1)$$

Для распределений, которые начинаются от $x = 0$, соответственно примем $a = 0$, а b получим из вышеприведенных рассуждений. Если границы интервалов не удается определить аналитически, то можно применить любые численные процедуры решения нелинейного уравнения. Для нахождения экстремума распределения также могут потребоваться численные методы, если это невозможно сделать аналитически.

Весь интервал разобьем на множество равных подынтервалов $[a = x_0, x_1, ..., x_k = b]$. Количество интервалов, равное k, будет задаваться пользователем. Внутри каждого подынтервала среднее значение функции плотности определяется по формуле

$$\frac{f(x_{j-1}) + f(x_j)}{2} = f(\tilde{x}_j), \quad (2)$$

где $\tilde{x}_j \in [x_{j-1}, x_j]$.

Вероятность попадания случайной величины X в *j*-й интервал есть площадь, ограниченная функцией плотности на отрезке $[x_{j-1}, x_j]$. Обозначим ее через p_j:

$$P(x_{j-1} \le X \le x_j) = p_j = \frac{f(x_{j-1}) + f(x_j)}{2} \cdot (x_j - x_{j-1}). \quad (3)$$

На основании закона больших чисел с ростом объема выборки значений случайной величины *n* частота наступления случайного события стремится к вероятности этого события. Тогда можно принять, что

$$p_j = \frac{n_j}{n}, \quad (4)$$

где n_j – количество попаданий значений случайной величины X в интервал $[x_{j-1}, x_j]$ из общего объема значений.

Определим количество точек n_j, которые нужно поместить в интервал $[x_{j-1}, x_j]$ для того, чтобы сгенерированная выборка соответствовала заданному закону распределения (например, равномерному закону или любой встроенной функции генерации конкретного языка программирования):

$$n_j = \left[p_j \cdot n \right]. \quad (5)$$

В данном случае качество встроенного генератора случайных чисел становится некритичным.

Совокупность точек, полученных описанным выше способом на всех интервалах, образует выборку случайной величины X, распределенной по заданному закону $f(x, m_x, \sigma_x, ...)$, и обладает всеми требуемыми значениями параметров. Выбор точек из представленного набора значений может производиться также любой встроенной функцией извлечения данных из массива значений.

Необходимо проверить, что полученная выборка чисел действительно имеет требуемый закон распределения и обладает нужными характеристиками, для чего по этой выборке строится гистограмма распределения, которая сравнивается с заказанной функцией плотности.

Приведем результаты численного моделирования предложенного алгоритма и оценки характеристик случайной величины по сгенерированной выборке.

Рассмотрим случайную величину, распределенную по логнормальному закону $X \sim \log N(\mu, \sigma^2)$, $\mu = 0, \sigma^2 = 0,5$ [4; 5]. Плотность и закон распределения логнормального закона имеют вид

$$f(X, \mu, \sigma) = \begin{cases} \dfrac{1}{\sqrt{2\pi} X \sigma} \exp\left(\dfrac{-(\ln X - \mu)^2}{2\sigma^2} \right), & X > 0, \\ 0, & X \le 0, \end{cases}$$

$$F(X, \mu, \sigma) = \int_{-\infty}^{X} f(t, \mu, \sigma) dt. \quad (6)$$

Оценка плотности распределения, приведенная на рис. 3, построена на основе выборки объемом $s = 100$, количество подынтервалов $k = 20$. Выборочные оценки параметров закона распределения, усредненные по 1 000 экспериментам, составили $\hat{\mu} = -0,002$, $\hat{\sigma}^2 = 0,49$.

График, приведенный на рис. 4, построен на основе выборки объемом $s = 500$, количество подынтервалов $k = 30$. Выборочные оценки параметров закона распределения, также усредненные по 1 000 экспериментам, составили $\hat{\mu} = -0,011$, $\hat{\sigma}^2 = 0,497$.

Далее рассмотрим случайную величиу, распределенную по двухпараметрическому закону Вейбулла [4; 6]. Этот закон и плотность его распределения имеют вид

$$F(X, a, b) = \begin{cases} 1 - e^{-\left(\frac{X}{b}\right)^a}, & X \ge 0, \\ 0, & X < 0, \end{cases}$$

$$f(X, a, b) = \begin{cases} \dfrac{a}{b} \left(\dfrac{X}{b} \right)^{a-1} e^{-\left(\frac{X}{b}\right)^a}, & X \ge 0, \\ 0, & X < 0, \end{cases} \quad (7)$$

где *a* – параметр формы; *b* – параметр масштаба.

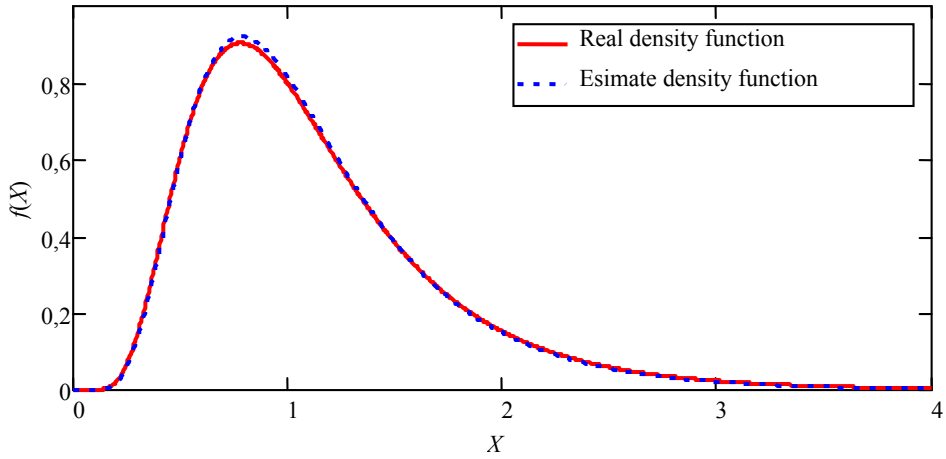

Рис. 3. Плотность распределения логнормального закона и ее оценка:
$S = 100$; $k = 20$

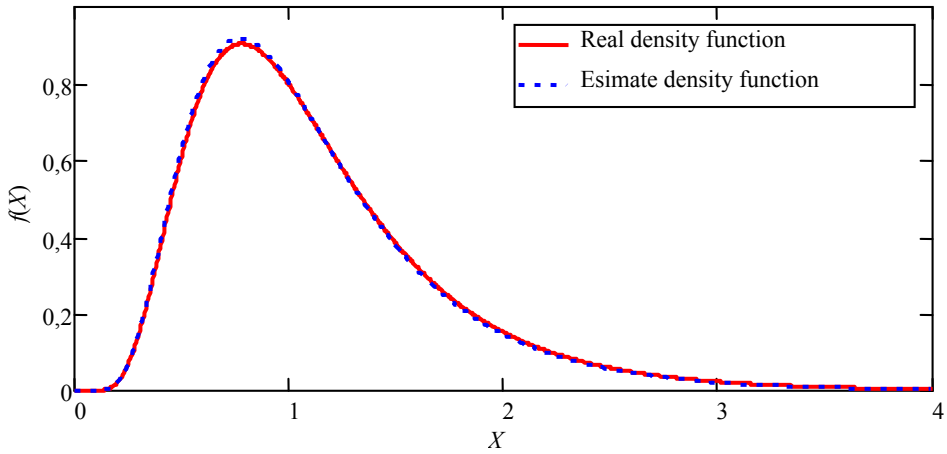

Рис. 4. Плотность распределения логнормального закона и ее оценка:
$S = 500$; $k = 30$

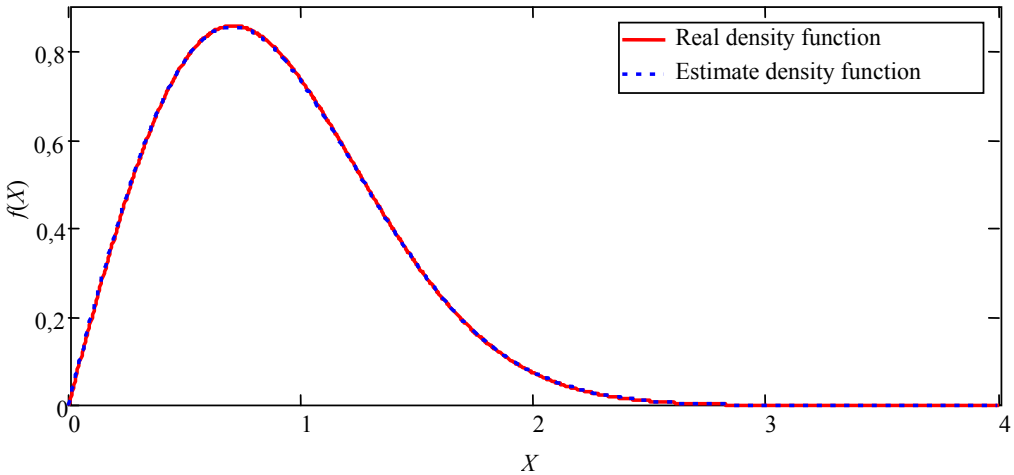

Рис. 5. Функция плотности распределения Вейбулла и ее оценка:
объем генерируемой выборки $s = 100$; количество подынтервалов $k = 12$; истинные значения параметров $a = 2$, $b = 1$; выборочные значения параметров $\hat{a} = 1,991$, $\hat{b} = 1$

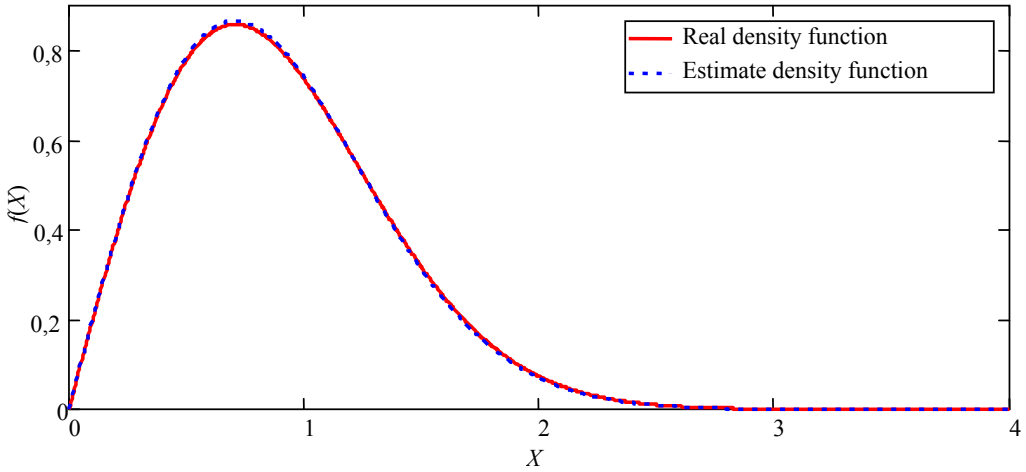

Рис. 6. Функция плотности распределения Вейбулла и ее оценка:
объем генерируемой выборки $s = 500$; количество подынтервалов $k = 30$; истинные значения параметров $a = 2$, $b = 1$; выборочные значения параметров $\hat{a} = 2{,}012$, $\hat{b} = 0{,}994$

В результате проведения экспериментов получены следующие графики (рис. 5, 6).

Построим гистограммы для логнормального распределения (рис. 7, 8) и распределения Вейбулла (рис. 9, 10).

Рис. 7. Гистограмма для логнормального закона:
параметры $\mu = 0$, $\sigma^2 = 0{,}5$; объем выборки $n = 100$,
количество подынтервалов $k = 20$

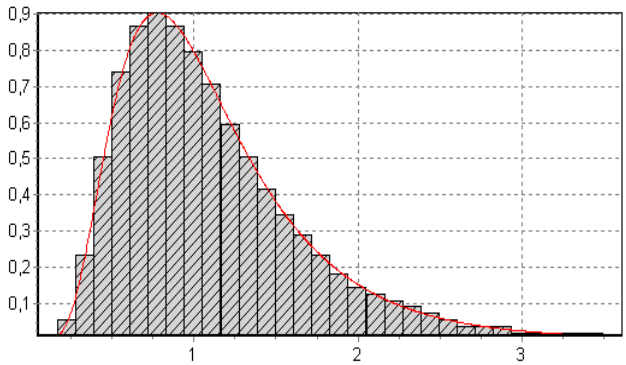

Рис. 8. Гистограмма для логнормального закона:
параметры $\mu = 0$, $\sigma^2 = 0{,}5$; объем выборки $n = 500$,
количество подынтервалов $k = 30$

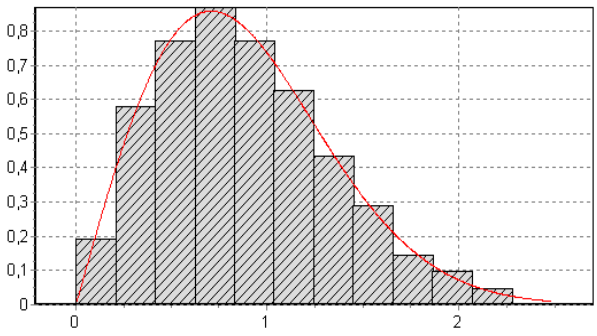

Рис. 9. Гистограмма закона Вейбулла
параметры $a = 2$, $b = 1$; объем выборки $n = 100$,
количество подынтервалов $k = 12$

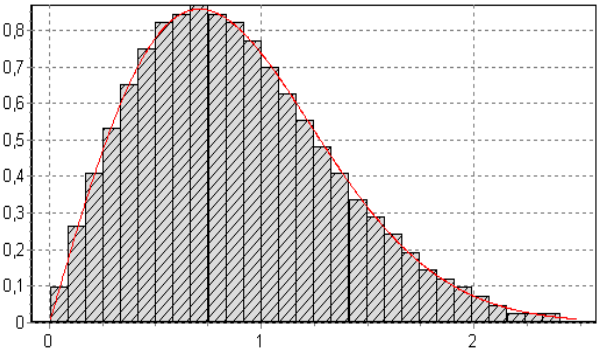

Рис. 10. Гистограмма закона Вейбулла:
параметры $a = 2$, $b = 1$; объем выборки $n = 500$,
количество подынтервалов $k = 30$

Дальнейшее развитие исследований касается построения генераторов для как можно большего количества известных функций распределения, изучения свойств сходимости каждого конкретного распределения, точности оценок числовых параметров распределения и случайной величины.

Библиографические ссылки

1. Медведев А. В. Теория непараметрических систем. Моделирование // Вестник СибГАУ. 2010. Вып. 4. С. 4–9.

2. Ермаков С. М., Михайлов Г. А. Курс статистического моделирования. М. : Наука, 1976.

3. Фадеева Л. Н. Математика для экономистов: Теория вероятностей и математическая статистика. М. : Эксмо, 2006.

4. Вероятность и математическая статистика : энциклопедия / под ред. Ю. В. Прохорова М. : Большая Рос. энцикл., 2003.

5. Первушин В. Ф., Сергеева Н. А. Генератор случайных чисел, распределенных по логнормальному закону // Решетневские чтения : материалы XIII Междунар. науч. конф. : в 2 ч. Ч. 2 / Сиб. гос. аэрокосмич. ун-т. Красноярск, 2009. С. 448–449.

6. Сергеева Н. А., Стрельников А. В. Генератор случайных чисел, распределенных по закону Вейбулла // Решетневские чтения : материалы XIII Междунар. науч. конф. : в 2 ч. Ч. 2 / Сиб. гос. аэрокосмич. ун-т. Красноярск, 2009. С. 451–453.

V. F. Pervushin, N. A. Sergeeva, A. V. Strelnikov

THE RANDOM SELECTION PRECISION GENERATOR

The algorithm of random value selection with adjusted distributions generation and numerical characteristics is considered (mathematical estimation, dispersion etc.).The algorithm proceedings of high precisions is showed on calculated examples of selection modeling. The selection generation common concept allows to extract this approach to the special category of random value generation algorithms. It was joined to the term "The Random selection precision generator".

Keywords: random value, retribution law, numeric characteristics, event probability, event frequency, selection size, bar chart.

УДК 519.6

В. Б. Бериков

АЛГОРИТМ АДАПТИВНОГО ПЛАНИРОВАНИЯ АНСАМБЛЯ ТАКСОНОМИЧЕСКИХ ДЕРЕВЬЕВ РЕШЕНИЙ*

Рассматривается подход к решению задач кластерного анализа, основанный на применении ансамбля таксономических деревьев решений. Предлагается алгоритм адаптивного планирования ансамбля, использующий расстояния между логическими высказываниями, описывающими кластеры. Приводятся результаты статистического моделирования, подтверждающие эффективность алгоритма.

Ключевые слова: кластерный анализ, ансамбль, дерево решений.

Одной из актуальных проблем кластерного анализа (таксономии, автоматической классификации «без учителя») является группировка объектов, описываемых разнотипными (количественными или качественными) переменными. Другая актуальная проблема связана с повышением устойчивости группировочных решений, так как в большинстве алгоритмов кластерного анализа результаты могут сильно меняться в зависимости от выбора начальных условий, порядка объектов, параметров работы алгоритма и т. п.

Наиболее перспективный подход к кластерному анализу при наличии разнотипных переменных основан на применении деревьев решений, которые позволяют получать легко интерпретируемую логическую модель группировки, выделять наиболее информативные факторы и не требуют задания метрики в разнотипном пространстве. Особенностью логико-вероятностного подхода, основанного на деревьях решений, является возможность не только разбивать заданное множество объектов на кластеры, но и строить иерархическое дерево, описывающее структуру разбиения.

Повысить устойчивость кластеризации можно с помощью ансамблей алгоритмов. При этом используются результаты группировки, полученные различными алгоритмами или одним алгоритмом, но с различными параметрами настройки, по различным подсистемам переменных и т. д. После построения ансамбля проводится нахождение итогового коллективного решения. Такой способ описан, например, в работе [1].

*Работа выполнена при финансовой поддержке Российского фонда фундаментальных исследований (проекты 08-07-00136а, 10-01-00113а, 09-07-12087-ofi_m) и Междисциплинарного интеграционного проекта Сибирского отделения Российской академии наук № 83.