СТАТИСТИЧЕСКОЕ МОДЕЛИРОВАНИЕ В ЧИСЛЕННОМ ЭКСПЕРИМЕНТЕ КЛАСТЕРИЗАЦИИ

П. Н. Звягин (Санкт-Петербург)

Имитационное моделирование – это метод, позволяющий строить модели, описывающие процессы так, как они проходили бы в действительности [1]. Такую модель можно «проиграть» во времени, как для одного испытания, так и заданного их множества. При этом результаты будут определяться случайным характером процессов. По этим данным можно получить достаточно устойчивую статистику.

Так, например, согласно [1], можно выделить две разновидности имитации, использующей строгие математические методы:

- метод Монте-Карло (статистические испытания);
- метод имитационного моделирования (статистическое моделирование).

В настоящей работе рассматривается задача отнесения поступающих данных при мониторинге ледовых нагрузок на корпус судна или неподвижной конструкции к нескольким группам. Эта задача, достаточно сложная при ее физической формулировке, может быть решена численно, с использованием в алгоритме подходов имитационного моделирования.

В качестве исходных данных были использованы наблюдения штатных тензостанций научно-экспедиционного судна «Академик Федоров», полученные во время летней экспедиции 2005 года к Северному полюсу. Тензостанции расположены в носовом и миделевом районах судна на высоте ледового пояса.

Предобработанные данные поступали на вход численного алгоритма, внутренние характеристики которого моделировались согласно предъявляемым требованиям.

Некоторые оптимизационные задачи формулируются таким образом, что их решение удобно находить при помощи итеративного численного алгоритма. Вид целевой функции и ограничений может быть таков, что можно связать переменные соотношениями, последовательное удовлетворение которым, с пересчетом значений переменных, приводит к оптимизации целевой функции.

Такому численному алгоритму дается старт из некоторой точки, как правило – случайной. При этом точка должна удовлетворять определенным в задаче условиям. Логично предположить, что, даже при наличии выраженного глобального оптимума, для его достижения из разных точек потребуется различное число итераций алгоритма. Кроме этого, при сильно нерегулярной поверхности целевой функции, из некоторых стартовых точек алгоритм может сходиться к локальным оптимумам. В таком случае целесообразным представляется проведение статистического эксперимента по старту алгоритма из различных точек. Такой подход имеет также некоторое отношение к методу Монте-Карло [2].

Рассмотрим метод кластеризации *C-means*, или, как его иногда называют в русскоязычной литературе, *C-средних*. В задаче, решаемой этим методом, требуется провести разбиение n имеющихся примеров $\bar{x}_j$, $j = 1,2,...,n$, на заданное число K групп, называемых также кластерами. Для этого, в результате решения задачи, необходимо найти значения элементов матрицы **U** принадлежности исходных примеров центрам кластеров $\bar{c}_i$ так, чтобы минимизировать функцию [3]

$$E = \sum_{i=1}^{K} \sum_{j=1}^{n} u_{ij}^{m} \cdot \left\| \bar{x}_j - \bar{c}_i \right\|^2 \to \min \qquad (1)$$

с учетом системы из n ограничений

$$\sum_{i=1}^{K} u_{ij} = 1,\ j = 1,2,...,n \qquad (2)$$

и ограничений на величину элементов u_{ij} матрицы **U**:

$$0 \le u_{ij} \le 1. \qquad (3)$$

Матрица принадлежности **U** имеет размерность $(K \times n)$. Векторы – центры кластеров $\bar{c}_i$, $i = 1..K$, имеют тот же размер, что и векторы исходных данных.

Изложенную задачу можно свести к минимизации функции Лагранжа

$$L = \sum_{i=1}^{K}\sum_{j=1}^{n} u_{ij}^{m} \cdot \left\| \bar{x}_j - \bar{c}_i \right\|^2 + \sum_{j=1}^{n} \lambda_j \left(1 - \sum_{i=1}^{K} u_{ij} \right) \to \min,\ 0 \le u_{ij} \le 1. \qquad (4)$$

Здесь переменные, по которым происходит оптимизация, – $\bar{c}_i$ и u_{ij}.

Итеративный алгоритм, получающийся в результате решении этой задачи, изложен в [3]. Алгоритм стартует, исходя из некоторой заданной матрицы **U**, удовлетворяющей начальным условиям (2) и (3).

Для каждого набора примеров $\bar{x}_j$, $j = 1,2,...,n$, поверхность, образуемая функцией Лагранжа *L* (4), имеет свой вид. Число примеров *n* может быть достаточно большим, также, как и их размерность. Это затрудняет анализ характера поверхности, образуемой *L*.

Поэтому целесообразным представляется использование стартовой матрицы **U**, инициированной случайным образом.

Охарактеризуем задание стартовой матрицы **U**.

Рассмотрим первый столбец u_{11}, u_{12}, ..., u_{1K}. Пусть первый элемент первого столбца u_{11} задан случайной величиной Z_1, распределенной на отрезке [0,1], чтобы удовлетворять условию (3). Тогда второй элемент u_{12} можно представить в виде реализации случайной величины Z_2 с распределением, зависящим от значения, которое приняла случайная величина Z_1, в силу условия (2). Распределение случайной величины Z_3, задающей третий элемент первого столбца матрицы **U**, будет зависеть от значений, принятых Z_2 и Z_1, и так далее. Последний элемент столбца находится как разность между единицей и всеми предыдущими элементами этого столбца.

Аналогичное рассуждение можно провести для всех остальных столбцов матрицы **U**.

Пусть все случайные величины, задающие элементы матрицы **U**, имеют равномерное распределение. Рассмотрим вновь элементы первого столбца. Если случайная величина Z_1, распределенная равномерно на отрезке [0,1], приняла значение z_1, то случайная величина Z_2 будет распределена равномерно на отрезке [0,1- z_1], чтобы удовлетворять условию (2). Если Z_2 приняла значение z_2 , то случайная величина Z_3 будет равномерно распределена на отрезке [0,1- z_1- z_2]. Последний элемент столбца матрицы **U**, u_{K1}, вычисляется как разность

$$u_{K1} = 1 - z_1 - z_2 - ... - z_{K-1}. \qquad (5)$$

Предположим, что нам необходимо промоделировать элементы матрицы **U** таким образом, чтобы начальная принадлежность ко всем кластерам была приблизительно одинаковой. В таком случае возможно применить моделирование случайных величин с задаваемой плотностью вероятности.

Рассмотрим такую плотность, что случайная величина, заданная при помощи нее, вероятнее примет значения, близкие к заданному центру M, чем далекие от него.

Пусть плотность задана на отрезке $[a,b]$ функцией $f(x)$, а на остальном множестве значений равна нулю. Функция распределения такой случайной величины выражается как

$$F(x)=\begin{cases} 0, & x \le a \\ \int_a^x f(t)dt, & a < x \le b \\ 1, & x > b \end{cases} \qquad (6)$$

Если известна функция распределения $F(x)$, то случайную величину возможно промоделировать, решив уравнение [4]:

$$F(x)=\alpha, \qquad (7)$$

где α – случайная величина, равномерно распределенная на отрезке [0,1].

Выгодно задавать такую функцию распределения $F(x)$, чтобы было удобно решать уравнение (7).

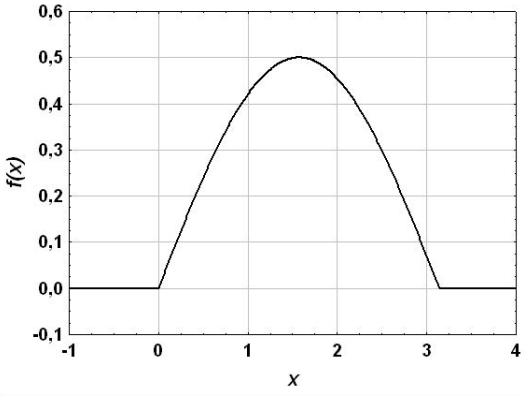

График плотности распределения (8)

Для моделирования значений принадлежности, подчиняющихся условию приблизительного равенства по всем кластерам, применялась случайная величина со следующей плотностью распределения (рисунок), реализации которой масштабировались в отрезок [0,2M]:

$$f(x)=\frac{1}{2}\sin x,\ 0 \le x \le \pi; \qquad f(x)=0, \qquad x<0,\ x>\pi. \qquad (8)$$

В этом случае реализацию случайной величины, распределенной по закону (8), можно получить, подставив вместо *y* в выражении (9) значение случайной величины, распределенной по равномерному закону на отрезке [0,1]:

$$x=\arccos(1-2y). \qquad (9)$$

Случайная величина *X*, реализации *x* которой получаются с использованием выражения (9), распределена на отрезке [0,π]. Для формирования элементов матрицы принадлежности **U** нас будут интересовать случайные величины, распределенные на отрезке [0,2M]:

$$M=\frac{1}{K}, \qquad (10)$$

где *K* – число кластеров. Таким образом, будет необходимо масштабировать получаемые по соотношению (9) реализации случайной величины *X*.

Для определенности рассмотрим формирование первого столбца матрицы **U**. Реализация *z* случайной величины Z, представляющая собой искомое значение элемента столбца матрицы **U**, будет выглядеть как

$$z = \frac{2M}{\pi} \cdot x. \qquad (11)$$

По формуле (11) можно найти все элементы столбца матрицы **U**, кроме последнего. Может так случиться, что сумма смоделированных первых *K-1* элементов столбца будет больше единицы. В таком случае следует повторять процесс моделирования элементов первого столбца заново, пока не удастся добиться суммы, меньшей единицы.

Последний элемент, чтобы удовлетворить ограничению (2), находится вычитанием из единицы всех смоделированных элементов этого столбца.

Введем параметр, позволяющий оценить отклонение значений матрицы **U** от величины M:

$$r = \sum_{i,j} \left(u_{ij} - M\right)^2. \qquad (12)$$

Так же можно рассматривать параметр, не зависящий от числа примеров и кластеров:

$$r_1 = \frac{\sum_{i,j}^{K,n} \left(u_{ij} - M\right)^2}{K \cdot n}. \qquad (13)$$

Проведем статистический эксперимент по составлению стартовой матрицы алгоритма *C-means* двумя способами. Первый способ будет использовать получение элементов как результат реализаций равномерно распределенных случайных величин. Второй способ будет использовать получение элементов как реализации случайных величин с заданным распределением по схеме (9)–(11), где *M* находится по соотношению (10). Проведем по 200 экспериментов для каждого способа, полученные наборы параметров *r* и r_1 представим в виде соответствующих выборок и найдем выборочные характеристики (табл. 1). Число столбцов матрицы *n* было принято равным 40.

Реализация статистического эксперимента проводилась при помощи специальной программы на языке C++, но для такого эксперимента возможно и использование различных математических пакетов [5].

Таблица 1

Статистики для параметров, характеризующих отклонение от значения *M* элементов стартовой матрицы U при различных способах задания матрицы

	Способ 1				***Способ 2***			
	r		r_1		*r*		r_1	
	Мат. ожидание	С.к.о.	Мат. ожидание	С.к.о.	Мат. ожидание	С.к.о.	Мат. ожидание	С.к.о.
K=3	8,8	1,09	0,073	0,009	3,07	0,47	0,026	0,0039
K=4	10,6	1,08	0,066	0,007	2,55	0,36	0,016	0,0022
K=5	12,0	1,16	0,06	0,006	2,14	0,3	0,011	0,0015
K=6	13,3	1,28	0,055	0,005	1,85	0,23	0,008	0,001
K=7	14,1	1,24	0,05	0,004	1,6	0,19	0,006	0,0007

По выборочным характеристикам, приведенным в табл. 1, видно, что при втором способе задания элементы матрицы **U** расположены существенно ближе к величине M, т. е. наблюдается меньшее математическое ожидание отклонения. Таким образом, второй способ задания вполне удовлетворяет предъявленному требованию.

Кроме того, в ряде случаев, при *использовании второго способа задания матрицы **U** сокращается число итераций алгоритма C*-means по переходу в точку, доставляющую минимум целевой функции. В качестве примера можно привести работу алгоритма на данных – 40 векторах, каждый длиной в 501 элемент, – обработанных результатах тензометрии корпуса судна при движении во льдах [6].

В табл. 2 представлено выборочное математическое ожидание и среднеквадратическое отклонение числа итераций по результатам 50 запусков алгоритма при разных способах задания стартовой матрицы **U**, при числе кластеров *K=3,* для данных по тензодатчику, расположенному в миделевой части судна.

Таблица 2

Статистики для числа итераций алгоритма *C-means* при первом и втором способах задания стартовой матрицы U для данных – периодограмм

Способ 1		***Способ 2***	
Мат. ожидание	С.к.о.	Мат. ожидание	С.к.о.
35,58	10,69	30,88	7,74

Вывод

В статье предложены имитационные подходы к формированию стартовой матрицы численного алгоритма *C-means* при помощи реализаций случайных величин с равномерным и заданным распределениями в задаче кластеризации результатов тензометрии корпуса продвигающегося во льдах судна.

Подход, использующий моделирование случайных величин с заданным распределением, позволяет, в ряде случаев, сократить время выполнения численного алгоритма. Кроме того, такой подход позволяет внести большую определенность в выбор области старта алгоритма, сохраняя при этом случайность стартового значения матрицы.

Литература

1. **Хемди А. Таха.** Введение в исследование операций. (Operations Research: An Introduction). М.: Вильямс, 2007.
2. **Вуколов Э. А.** Основы статистического анализа. М.: Форум-Инфра-М, 2004.
3. **Jang J. S.,Sun C. T., Mizutani E.** Neuro-fuzzy and soft computing. NY: Prentice Hall, 1997.
4. **Сушков Ю. А.** Статистические модели систем. СПб.: Изд-во СПбГУ, 2004.
5. **Звягин П. Н., Звягин К. Н.** Компьютерное моделирование нормально распределенных случайных величин//Сб. докладов конференции ИММОД-2005. СПб., 2005. С. 196–200.
6. **Звягин П. Н., Нечаев Ю. И., Тимофеев О. Я.** Применение аппарата нечеткой логики в системах мониторинга ледового воздействия//Труды НТК «Проблемы мореходных качеств судов и корабельной гидромеханики. XLII Крыловские чтения». СПб.: ЦНИИ им. академика А. Н. Крылова, 2006. С. 99–101.