Процедура построения кривых плотностей распределения Пирсона для многомодальных выборок

Борчик Е.М.; Башаримов В.В.; Якимов А.И.
Кафедра АСУ, электротехнический факультет
ГУВПО «Белорусско-Российский университет»
г. Могилев, Беларусь
e-mail: ykm@tut.by

***Аннотация* — Рассматривается задача построения статистических моделей распределений многомодальных выборок данных. Предложена процедура разделения исходной многомодальной выборки методами кластерного анализа на несколько однородных выборок-кластеров с последующим построением на каждой из них своей функции плотности обобщенного распределения Пирсона. Для проверки соответствия построенных статистических моделей распределений применяются статистические критери Пирсона, Колмогорова-Смирнова, Мизеса.**

Ключевые слова: обобщенное распределение Пирсона; статистические критерии проверки гипотез; методы кластерного анализа.

I. Введение

Пусть в ходе имитационных экспериментов получена выборка $X = \{x_i \mid x_i \in R, i = 1, \ldots, n\}$. Необходимо построить статистическую модель распределения выборочных данных (кривую), наилучшим образом описывающую данную выборку на исследуемом интервале $[a, b]$. В общем случае выборка X является многомодальной.

Поставленную задачу можно решить с использованием существующих «известных» законов распределений [1], рядов специального вида [2], семейств универсальных статистических моделей распределений [1, 2, 3].

Каждый из подходов имеет свои достоинства и недостатки. Например, особенностью семейств универсальных моделей распределений является возможность аппроксимации лишь одномодальных и U-образных распределений.

Многомодальность распределения указывает на неоднородность исследуемой выборки X. В этом случае предлагается разделение методами кластерного анализа [4] исходной выборки на несколько однородных выборок (кластеров) с последующим построением на каждой из них своей функции плотности распределения.

Построение функции плотности распределения семейства Пирсона, соответствующей эмпирическому распределению выборки X, реализовано в программном модуле «BelSim2#.random» [5, 6].

Процедура разделения исходной выборки X на кластеры, реализована в программно-технологическом комплексе имитации сложных систем BelSim2 [4, 7].

II. Построение функции плотности распределения семейства Пирсона

Для работы модуля «BelSim2#.random» рассчитываются необходимые точечные оценки выборки X, в частности, начальный момент первого порядка ν_1; центральные моменты $\mu_0, \ldots, \mu_4$; коэффициенты асимметрии и эксцесса, введённые Пирсоном [3]:

$$\beta_1 = \mu_3^2 / \mu_2^3, \quad \beta_2 = \mu_4 / \mu_2^2. \qquad (1)$$

Построение кривой $f(x)$ в соответствии с классификацией Пирсона выполняется в следующей последовательности.

Шаг В.1. Классификация типа кривой $f(x)$.

Проводится в соответствии со значениями коэффициентов β_1, β_2 вида (1) и показателя Пирсона классификации плотности распределения кривой [6]:

$$\aleph = \left(\beta_1(\beta_2+3)^2\right) \big/ \left(4(2\beta_2 - 3\beta_1 - 6)(4\beta_2 - 3\beta_1)\right). \qquad (2)$$

Шаг В.2. Оценка значений параметров $f(x)$.

Проводится в соответствии классическим методом моментов для функции плотности распределения $f(x)$, заданного на *Шаге В.1*. В соответствии с методикой, предложенной Пирсоном [2, 3], для построенной кривой $f(x)$ определяется нормирующий множитель N. Расчет N производится исходя из условия равенства единице интеграла от $f(x)$ на интервале $[a, b]$, в пределах которого производится построение кривой.

Шаг В.3. Поверка соответствия $f(x)$ выборке X.

Для проверки гипотезы о распределении выборки X по закону распределения с плотностью $f(x)$ выбраны критерии согласия Пирсона (χ^2), Колмогорова-Смирнова (λ), Мизеса (ω^2) [1].

Первыми применяются критерии χ^2 и λ. Если логические значения результатов их работы эквивалентны, то на этом этап статистической проверки гипотезы заканчивается. Иначе – дополнительно применяется критерий ω^2, результат работы которого принимается в качестве заключения о проверке гипотезы.

Сохраняемые результаты работы критериев χ^2, λ, ω^2: наблюдаемые значения критериев N_χ, N_λ, N_ω; критические значения критериев K_χ, K_λ, K_ω; отношение наблюдаемых значений критериев к критическим $dL_1 = N_\chi / K_\chi$, $dL_2 = N_\lambda / K_\lambda$, $dL_3 = N_\omega / K_\omega$; логические результаты работы критериев $bL_k = iif(dL_k < 1, true, false)$, $k = 1, 2, 3$ соответственно.

III. Определение статистической модели распределения выборочных данных $f^*(x)$, наилучшим образом описывающей выборку X

В случае отклонения статистическими критериями гипотезы о принадлежности выборки X закону распределения с плотностью $f(x)$, или по специальному запросу пользователя, производится построение с последующей проверкой комплексом статистических критериев χ^2, λ, ω^2 всех основных типов кривых семейства за исключением функции $f(x)$ проверенной на предыдущем этапе.

По запросу пользователя производится построение функций распределений трёх частных случаев распределений: равномерного, нормального и экспоненциального с последующей проверкой комплексом статистических критериев χ^2, λ, ω^2.

Выбор наилучшей статистической модели производится на множестве тех кривых $\{f_i(x)\}$, которые не отклонены статистическими критериями.

Шаг E.1. В том случае, если не для всех функций $f_i(x)$, $i = 1, ..., |f_i(x)|$, применено одинаковое количество критериев $nk \in \{2, 3\}$, для возможности определения $f^*(x)$ проводится дополнительный расчёт критерия ω^2.

*Шаг E.*2. Функция $f^*(x) = f_{i_0}(x) \in \{f_i(x)\}$ описывает выборку X наилучшим образом на интервале $[a, b]$ тогда и только тогда, когда выполняется условие:

$$\rho(X, f_{i_0}(x)) = \min\{\rho(X, f_i(x)) | \, i = 1, ..., |f_i(x)|\}, \quad (3)$$

где $\rho(X, f_i(x))$ – евклидова метрика от отношений dL_k:

$$\rho(X, f_i(x)) = \sqrt{\sum_{k=1}^{nk} (dL_k(X, f_i(x)))^2}, \quad (4)$$

$nk \in \{2,3\}$ – количество применённых к $f_i(x)$ критериев.

В том случае, если все построенные кривые отклонены статистическими критериями, нет возможности построения кривой плотности распределения семейства Пирсона, описывающей выборку X на исследуемом интервале $[a, b]$.

IV. Кластерный анализ данных

Кластеризация данных является одним из этапов построения законченного аналитического решения поставленной задачи. Часто легче выделить группы схожих объектов, изучить их особенности и построить для каждой группы отдельную модель, чем создавать одну общую модель для всех данных.

Шаг F.1. Производится разделение исходной выборки X на кластеры K_j, $j = 1, ..., |K_j|$, $X = K_1 \cup ... \cup K_{|K_j|}$.

Количество кластеров $|K_j|$ определяется в зависимости от количества областей отдельных «пиков» в полигоне частот исследуемой выборки X.

Для кластеризации выборки выбраны методы: Tree Clustering, K-Means, Fuzzy Relation Clustering [4].

Обобщение результатов кластеризации выборки X несколькими методами кластерного анализа производится в соответствии с *Утверждением* 1.

Утверждение 1. Элемент $x_i \in X$ является элементом кластера K_j, $j \in 1, ..., |K_j|$, тогда и только тогда, когда x_i отнесён к данному кластеру, по крайней мере, двумя применяемыми методами кластерного анализа из трёх.

Шаг F.2. В том случае, если выявлены выбросы, производится их сглаживание. Повторяется *Шаг F.1*.

Шаг F.3. Для каждого из кластеров K_j, $j = 1, ..., |K_j|$, производится построение функции плотности распределения $f_j^*(x)$, соответствующей эмпирическому распределению элементов данного кластера.

В случае, если для кластера K_{j_0}, $j_0 \in 1, ..., |K_j|$ все построенные кривые отклонены статистическими критериями, для K_{j_0} нет возможности построения кривой $f_{j_0}^*(x)$. Рекомендуется разделение K_{j_0} на несколько кластеров меньшего размера (шаг *F.1.*) с последующим построением на каждой из них своей функции плотности распределения.

[1] Большев, Л.Н. Таблицы математической статистики / Л.Н Большев, Н.В. Смирнов. — М.: Наука, 1983. – 416 с.

[2] Кендалл, М. Теория распределений: пер. с англ. / М. Кендалл. А. Стюарт. – М.: Наука, 1966. – 588 с.

[3] Elderton, W. P. Frequency curves and correlation / W. P. Elderton, 4 ed., Camb., 1953. – 172 с.

[4] Якимов, А. И. О совместном использовании методов кластерного анализа многомерных данных / А.И. Якимов, Е.М. Борчик, В.В. Башаримов // Доклады БГУИР, в печати.

[5] Программный модуль расчета кривой плотности распределения случайной составляющей в последовательности данных «BelSim2#.random» : свидетельство о регистрации компьютерной программы № 306 / Е. А. Якимов, А. А. Ковалевич, Е. М. Борчик, В. В. Башаримов. – Минск: НЦИС, 2011. – Заявка № С20110024. – Дата подачи: 04.04.2011.

[6] Якимов, А. И. Методика построения кривой семейства плотностей обобщённого распределения Пирсона для исследуемой выборки / А. И. Якимов, Е. М. Борчик, В. В. Башаримов // Системный анализ и информационные технологии: материалы междунар. науч.-техн. конф. SAIT 2011 (23–28 мая 2011 г., Киев, Украина) – К.: НТУУ «КПИ», 2011. – С 347.

[7] Якимов, А. И. Технология имитационного моделирования систем управления промышленных предприятий : монография / А. И. Якимов. – Могилев: Белорус.-Рос. ун-т, 2010. – 304 с.: ил.