УДК 519.234

А.И. Арифов, А.Е. Архипов

О НЕПАРАМЕТРИЧЕСКОМ ПОДХОДЕ К ИДЕНТИФИКАЦИИ АНОМАЛЬНЫХ ДАННЫХ В ВЫБОРКАХ ОДНОВЕРШИННЫХ РАСПРЕДЕЛЕНИЙ

Введение

В литературе [1,2] описано достаточно много методов идентификации аномальных данных (АД) или выбросов из основной группы исходных данных. Однако на практике они мало применимы, поскольку имеют целый ряд значительных ограничений, таких как наличие сведений о виде закона распределения случайной величины (СВ), знание параметров выборки (мат. ожидание, дисперсия), количество присутствующих аномалий (кратность), число доступных для анализа элементов выборки. Поэтому представляется актуальным поиск непараметрических, достаточно общих (работающих для разных распределений) методов идентификации АД, по возможности лишенных вышеперечисленных недостатков.

Постановка задачи

Предельные (пороговые) значения Δ, превышение которых полученными в эксперименте данными позволяет относить их к АД, можно найти из неравенства Чебышева

$$P\{|X - m_x| \geqslant \Delta\} \leqslant D_x/\Delta^2.$$

Например, при $\Delta = 3\sigma_x$ $P\{|X - m_x| \geqslant 3\sigma_x\} \leqslant \frac{1}{9} = 0.1(1)$.

На практике для модельных законов распределения эта вероятность меньше. Так для нормального закона она равна ≈ 0.003, для равномерного – нулю.

Более подробно значения вероятностей $P\{|X - m_x| \geqslant \Delta\}$ для различных видов распределений (неизвестного, Лапласа, нормального, равномерного и Симпсона) при заданной величине порога Δ приведены в таблице 1.

Очевидно, что при одном и том же значении Δ для разных распределений будем получать разные вероятности $P\{\ldots\}$, т.е. задавать пороговое значение Δ для принятия решения о принадлежности результата к АД нельзя, аналогично нельзя использовать нормированное значение Δ/σ_x. Смысл имеет задание предельных значений $P_{\text{пред}}$, близость которых к 0 гарантирует тот или иной уровень надежности разделения исходного массива данных на кондиционные и аномальные. Можно ввести набор предельных значений

 ISSN 1562-9945

Таблица 1

Вероятность превышения порога Δ, $P\{|X - m_x| \geqslant \Delta\}$

Δ	σ_x^2/Δ^2	Вид распределения			
		Лапласа	Нормальное	Равномерное	Симпсона
$1.5\sigma_x$	0.44445	0.11987	0.13361	0.56667	0.15026
$2.0\sigma_x$	0.25000	0.05911	0.02275	0.43365	0.03367
$3.0\sigma_x$	0.1111(1)	0.01440	0.00270	0.13397	0.00000
$4.0\sigma_x$	0.06250	0.00349	0.00006	0.00000	0.00000

$$P_{\text{пред1}} < P_{\text{пред2}} < \ldots < P_{\text{предк}},$$

характеризующих различную степень надежности принятого решения в смысле ошибок I и II рода, в зависимости от смысла задачи (что более критично: пропуск АД или ошибочное отсеивание кондиционных).

От чего зависит значение Δ при выбранном конкретном значении $P_{\text{пред}}$? Очевидно, от характеристик формы распределения плотности вероятности $f(x)$, т.е. $\Delta = \varphi(f(x))$, следовательно, по своей сути, Δ- функционал.

За редким исключением (равномерный закон, закон Симпсона, законы, производимые от равномерного путем суммирования конечного числа одинаково распределенных СВ) модельные законы имеют “хвосты”, уходящие на бесконечность. В реальных ситуациях когда объект исследования – параметр, имеющий конкретное конечное значение, а измерительная система характеризуется конечной шкалой, т.е. и ошибка измерения конечна, трудно полагать, что модельные распределения адекватны реальным в области возможных АД (именно из-за бесконечных “хвостов” модельных распределений).

Поэтому нужно либо строить свои модели, имеющие конечные (ограниченные) “хвосты”, либо использовать генераторы значений СВ, имеющие ограниченные шкалы. Последние позволяют эмпирически определить значения Δ_j, соответствующие заданным значениям $P_{\text{пред}j}, j = 1, \ldots, k$. Определение Δ_j не должно быть зависимым от вида закона (модели) распределения данных, т.е. при расчете Δ_j должны учитываться лишь какие-то тенденции в форме закона распределения, особенно в маловероятных областях (в “хвостах”).

В реальной ситуации актуальна следующая задача. Пусть по исходным данным построена гистограмма для $f(x)$. Хотелось бы, найдя робастную оценку $\widetilde{m_x}$, “продвигаясь” в расчетах от нее влево и вправо, получить наборы выборочных характеристик формы исходного неизвестного распределения $f(x)$, по которым можно было бы прогнозировать значения $\Delta_{j\text{левое}}$ и $\Delta_{j\text{правое}}$ для заданных предельных значений $P_{\text{пред}j}$.

Сделать это можно было бы, рассчитывая для левого “хвоста” выборочные моменты $\lambda_{-1}, \lambda_{-2}, \lambda_{-3}, \ldots$, используемые для предсказания $\widetilde{\Delta}_{\text{левое}} = \lambda_{\text{левое}}(\lambda_{-1}, \lambda_{-2}, \lambda_{-3}, \ldots)$, представляющего собой некоторую оценку зна-

чения $\Delta_{j\text{левое}} = \varphi(f(x))$, а для “правого” хвоста – $\lambda_1 = \upsilon(\widetilde{f}(x)), \lambda_2 = \upsilon(\widetilde{f}(x)), \ldots$, по которым прогнозируется $\widetilde{\Delta}_{\text{правое}} = \lambda_{\text{правое}}(\lambda_1, \lambda_2, \lambda_3, \ldots)$.

Расчет моментов λ_{-i}, λ_i должен производиться по нарастающей влево и вправо от $\widetilde{m_x}$ выборки исходных данных, т.е. не использовать возможных аномальных значений, располагающихся крайне слева (справа).

В качестве классов $f(x)$ целесообразно использовать симметричные остроконечные (лапласоподобные) и шапкообразные (гауссоподобные) распределения с переводом последних исследований в область несимметричных форм.

Возможно, что для симметричных, для увеличения крутизны формы, целесообразно “складывать” левую и правую половины распределений, что увеличит число данных, участвующих в расчете прогнозирующей последовательности $\lambda_1, \lambda_2, \lambda_3, \ldots$ и повысит точность прогноза $\Delta_{правое} = |\Delta_{левое}|$.

Данные, используемые уже непосредственно для построения прогноза (т.е. не исходные данные) должны быть свободны от размерности хвостов. Приведение вида $x' = x/\sigma_x$ делать нельзя, т.к σ_x неизвестно. Проще нумеровать интервалы варьирования, используемые при построении гистограммы: $\delta_1 = x_1/x_1 = 1, \delta_2 = x_2/x_1 = 2, \ldots, \delta_j = x_j/x_1 = j$.

В качестве ординат целесообразно использовать относительные частоты $\nu_i = \frac{n_i}{N}$, где n_i – абсолютная частота попадания в i-й интервал варьирования, $N = \sum\limits_i n_i$. В этом случае ординаты не зависят от объема N выборки.

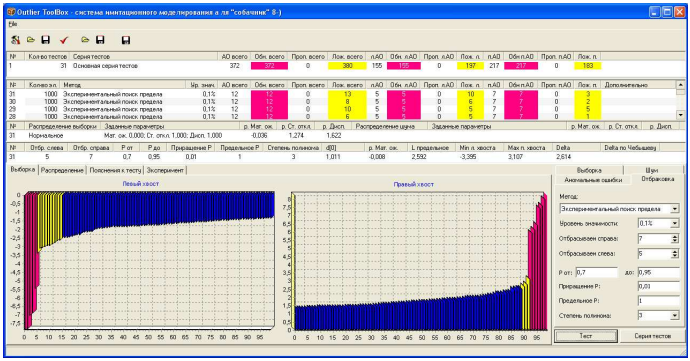
При таком формате входных данных (безразмерные значения и по абсциссе и по ординате) исследуется и учитывается только форма распределения $\widetilde{f}(x)$, без необходимости учета масштаба. Получение именованных значений по оси абсцисс достигается умножением безразмерного значения на x_1.

Для получения значения $\widetilde{\Delta}$ можно использовать различные прогнозные модели [3, 4]. В простейшем случае они могут быть представлены моделью авторегрессии, либо интегрированной авторегрессии, где прогноз будет оценкой накопления относительной частости, непосредственно определяющей значение $1 - P_{пред}$.

Реализация

Описанный выше подход к обнаружению АД опробован на основе имитационной системы Outlier Tool Box, основное назначение которой – определение области эффективности методов идентификации АД на наборе сгенерированных тестовых данных. На данный момент метод тестировался на выборках двойного равномерного (Симпсона), Гаусса и Лапласа законов распределения от 100 до 100000 точек. Независимо от типов распределения показал довольно высокую точность отбраковки, обнаружив только в части тестов ошибки второго рода – т.е. наряду с правильно идентифицированными АД присутствовали и ложно обнаруженные.

На рис. 1 показан фрагмент тестирования метода на выборках нормального распределения в 1000 точек. Хорошо видно, что на каждом хвосте наряду с корректно отбракованными АД (максимальные всплески), присутствуют и ложно обнаруженные (светлые) элементы. Однако их количество – в данном случае слева 10 и справа 3 незначительно для выборки в 1000 элементов.



На данный момент с определенностью можно сказать следующее. Метод устойчиво определяет область вероятных аномалий, куда вместе с АД входит незначительное количество кондиционных данных.

Резервы подхода еще далеко не исчерпаны. Основные направления оптимизации: подбор модели прогноза $\widetilde{\Delta}$, выбор $P_{пред}$, тестирование на широком спектре чистых и смешанных одновершинных распределений, повышение устойчивости работы.

Выводы

Так как неизвестно, насколько модельное распределение в области маловероятных значений (“хвосты” распределений) соответствуют распределениям случайных величин, встречающихся в практике обработки, воспользоваться результатами исследования модельных распределений нельзя, нет гарантий соответствия (адекватности) модели реальным данным в областях малых значений плотности вероятности. Скорее всего, их и быть не может из-за стремления “хвостов” модельных распределений к ∞, а реальных – к некоторому конечному значению.

Поэтому правила отбраковки АД следует формировать в ходе анализа имеющихся исходных данных, по результатам анализа получать Δ, т.е. информацию о форме распределения брать непосредственно из выборки, тогда это корректно.

Практика имитационного моделирования подтверждает верность предложенного подхода.

Литература

1. Айвазян С.А. и др. Прикладная статистика: Основы моделирования и первичная обработка данных. Справочное изд. /С.А. Айвазян,

И.С. Енюков, Л.Д. Мешалкин. М.: Финансы и статистика, 1983. - 471 с.

2. Основы испытаний летательных аппаратов: Учебник для вузов / Е.И. Криницкий, Л.Н. Александровская, В.С. Мельников, Н.А. Максимов; Под общей редакцией Е.И. Криницкого, - М.: Машиностроение, 1989. - 312 с. ил.
3. Химмельблау Д. Анализ процессов статистическими методами. – М.: Мир, 1973. -304 с.
4. Клейтен Дж. Статистические методы в имитационном моделировании. В 2-х томах. – М.: Статистика, 1978. – 507 с.

Получено 23.05.2008

 ISSN 1562-9945