Нами показана истинность положений П1, П2, П3, на основании которых формируется для произвольного шага алгоритма поиска полуоткрытый интервал неопределенности относительно точки с характерным признаком и выбирается одна из возможных стратегий: оптимистическая или пессимистическая.

Этим самым достигается оптимальность (см. соотношение (3) [3]) алгоритмов поиска, помехоустойчивых к тем или иным нерегулярным несимметричным виртуальным последовательностям положительной полярности. Цифровые автоматы, функционирование которых задают ориентированные графы переходов таких алгоритмов поиска, будут формировать за меньшее количество шагов псевдослучайную подстановку и тем самым будут уменьшать количество символов в шифротексте.

ПЕРЕЧЕНЬ ССЫЛОК

1. *Алипов Н. В.* Дискретные автоматы с псевдослучайными переходами и подстановочные методы защиты информации на их основе / Алипов Н. В. // Радиоэлектроника и информатика. – 2001. – № 4. – С. 95–98.
2. *Алипов Н. В.* Структура цифрового автомата с псевдослучайными переходами из начального состояния в одно и тоже конечное состояние / Алипов Н. В., Кораблев Н. М., Хиль М. И., Гусятин М. В. // Радіоелектроніка. Інформатика. Управління. – 2006. – № 2. – С. 102–109.
3. *Алипов Н. В.* Примеры построения ориентированных графов переходов цифрового автомата с псевдослучайными переходами / Алипов Н. В., Кораблев Н. М., Хиль М. И., Гусятин М. В. // Радіоелектроніка. Інформатика. Управління. – 2007. – № 1. – С. 97–105.
4. *Алипов Н. В.* Синтез оптимальных полихотомичных опросников для угадывания числа с ложными ответами / Алипов Н. В. // Проблемы бионики. – 1987. – Вып. 38. – С. 108–117.

Надійшла 3.09.2008

Визначено співвідношення ефективності застосування песимістичної й оптимістичної стратегії на j-му кроці алгоритму, на підставі яких обирають крок застосування песимістичної стратегії. На конкретних прикладах проілюстровані характерні випадки вибору стратегій пошуку.

The correlations of efficiency application of pessimistic and optimistic strat-egy are determined on a j-step algorithm, on the basis of which the step of pessimistic strategy application is selected. The concrete examples demon-strate the typical cases of strategies' search selection.

УДК 519.234

А. Е. Архипов, А. И. Арифов

ИСПОЛЬЗОВАНИЕ ПРОГНОЗНОГО ПОДХОДА ДЛЯ НЕПАРАМЕТРИЧЕСКОГО ОБНАРУЖЕНИЯ АНОМАЛЬНЫХ ДАННЫХ В ВЫБОРКАХ ОДНОВЕРШИННЫХ РАСПРЕДЕЛЕНИЙ

Предлагается непараметрический подход к выявлению аномальных данных, базирующийся на прогнозном определении границы области достоверных значений. Приведен ряд методов, позволяющий реализовать изложенный подход на практике.

ВВЕДЕНИЕ

Практическое решение прикладных задач с использованием экспериментально полученных данных (в частности, результатов измерений) показывает, что в общей совокупности исходных данных встречаются отдельные результаты, значения которых резко отличаются от остальных. Эти результаты получили название аномальных данных (АД) (другие названия: аномальные результаты измерений [1], грубые ошибки [2], резко выделяющиеся значения [3, 4], «подозрительные», «загрязняющие» значения [4], аномальные погрешности или ошибки [5]). К сожалению, достаточно строгое определение термина аномальные данные оказалось сложной задачей. Встречающиеся в литературе выражения вида: «ненормально большие погрешности (типа промах)» [5], «результаты наблюдений, которые сильно отличаются от центра распределения» [3] или подобные им не дают достаточно полного и адекватного представления об особенностях и свойствах АД.

Более удачным представляется описание АД так называемыми смесевыми моделями [6], из которых одной из первых и наиболее известных является модель Тьюки «засоренного» нормального распределения. Согласно этой модели, элементы исходной совокупности данных «извлекаются» из генеральной совокупности, заданной функцией плотности вероятности вида

$$f(z) = (1-\gamma)\varphi(z, \mu, \sigma_z^2) + \gamma\varphi(z, \mu\sigma_\alpha^2), \qquad (1)$$

где $\varphi(z, \mu, \sigma^2)$ – плотность нормального распределения со средним (математическим ожиданием) μ и

дисперсией σ^2, γ – вероятность появления АД, σ_z^2 – дисперсия достоверных измерений, σ_α^2 – дисперсия «засоряющей» совокупности. Обычно $\sigma_z^2 \ll \sigma_\alpha^2$, а $\gamma \ll 1$ (по некоторым оценкам $\gamma = 10^{-5} - 10^{-8}$).

Дальнейшим развитием модели Тьюки являются модели Хьюбера и Шурыгина [2, 6]. Последняя описывается выражением вида

$$f(z) = (1-\gamma)\psi(z, \mu, \sigma_z^2) + \gamma h(z-\mu, \sigma_\alpha^2), \qquad (2)$$

где основное распределение $\psi(z, \mu, \sigma_z^2)$ задается произвольной функцией плотности вероятности, а «засоряющее» распределение $h(z-\mu, \sigma_\alpha^2)$ – некоторое, обычно симметричное распределение.

Смесевые модели дают достаточно адекватное формализованное описание реальных данных, содержащих АД, и позволяют выделить перспективные подходы для выявления АД, в частности, приводящие к методам статистической классификации данных. Так, если полагать, что элементы исходной совокупности данных могут принадлежать только к одному из двух классов (достоверных данных или АД), приходим к классической задаче обнаружения (частный случай задачи распознавания при числе классов, равном 2). Однако при практическом применении методов статистической классификации к обнаружению АД сталкиваемся с рядом трудностей.

Во-первых, появление АД относится к ряду так называемых редких событий, для которых попытки оценить распределение $h(z-\mu, \sigma_\alpha^2)$ оказываются безуспешными ввиду недостаточного объема выборки АД (из-за крайне низкой вероятности γ). Поэтому обычно исследователь вынужден совершенно произвольно задавать как вид, так и параметры $h(z-\mu, \sigma_\alpha^2)$.

Во-вторых, периферийные области («хвосты») основного распределения $\psi(z, \mu, \sigma_z^2)$ формируются по результатам наблюдений, имеющих низкую вероятность появления. Поэтому подбор модели распределения $\psi(z, \mu, \sigma_z^2)$ в основном осуществляется исходя из соображений близости центральных частей модельного и эмпирического распределений. Поэтому близость реального и модельного распределений в периферийных областях весьма условна.

Из сделанных выше двух замечаний следует, что сведения о форме распределений в области, в которой должна лежать граница, отделяющая достоверные результаты от АД, фактически отсутствуют, предполагаемое строго решение задачи обнаружения АД по сути является производным и зависящим от опыта эксперта, осуществляющего обработку данных, а применяемый статистический подход и сопровождающие его математические выкладки лишь создают видимость математической строгости и точности.

ПОСТАНОВКА ЗАДАЧИ

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).

Предположение об ограниченности области существования достоверных значений z позволяет реализовать подход к выявлению АД, состоящий в оценке левого $z_л$ и правого $z_п$ граничных значений этой области и последующего обнаружения АД применением системы пороговых соотношений

$$\begin{cases} z_п \le z_i \le z_л, \quad z_i \in Z, \quad i = \overline{1, n}, \\ z_i < z_л, \quad z_i > z_п, \quad z_i \in \alpha, \quad i = \overline{1, n}, \end{cases} \qquad (3)$$

где z_i – проверяемый элемент совокупности $\{z_i\}$ исходных данных, Z – генеральная совокупность достоверных данных, α – генеральная совокупность АД.

Таким образом, основная проблема в реализации изложенного подхода к обнаружению АД состоит в оценивании граничных значений $z_л$, $z_п$, решение которой рассматривается ниже.

ВЫЯВЛЕНИЕ АНОМАЛЬНЫХ ДАННЫХ МЕТОДОМ ПРОГНОЗА ГРАНИЧНЫХ ЗНАЧЕНИЙ

Пусть эмпирическая функция распределения исходных данных представлена гистограммными оценками, для которых известна совокупность относительных частот $w_i = n_i/n$ попадания в интервал варьирования, рассчитанных для последовательности равных интервалов варьирования длиной Δ_z, где n_i – абсолютная частота попадания наблюдаемых значений в i-й интервал. Рассмотрим задачу оценивания только правого граничного значения z_n, что не ограничивает общности получаемых решений.

Предположим, процентное содержание АД среди положительных значений исходной выборки $\{z_i\}$, $i = \overline{1, n}$ не превышает β % от ее общего объема n, т. е. возможное число этих АД равняется $n_\alpha = \mathrm{ent}[n\beta/100]$, где ent[.] – операция выделения целой части числа. Отделим от исходной выборки n_α ее наибольших элементов и по оставшимся построим гистограмму интервального вариационного ряда. При таком представлении исходных выборочных данных оценивание граничного значения z_n можно реализовать, применяя методы прогноза временных рядов.

Действительно, последовательность середин интервалов варьирования, расположенных справа от значения $z = 0$, образует регулярную последовательность $\{z^{(1)}, z^{(2)}, z^{(3)}, \ldots\} = \{z^{(j)}\} = \{\Delta_z/2, 3\Delta_z/2, 5\Delta_z/2, \ldots\}$, являющуюся аналогом временной координаты временного ряда, а собственно «значениями временного ряда» являются относительные частоты соответствующих интервалов варьирования: $\{w_j\} = \{w_1, w_2, w_3, \ldots\}$ (рис. 1).

Выполняя оценивание значения z_n, необходимо учитывать условие, вытекающее из самой постановки задачи:

$$w(z_n) = 0, \qquad (4)$$

т. е. интерес представляет не собственно прогноз некоторого произвольного «значения временного ряда», а именно то значение координаты z, при котором выполняется условие (4). Эта особенность решаемой задачи в ряде случаев существенно упрощает получение решения. Например, если предположить, что ряду $\{w_j(z^{(i)})\}$ соответствует линейный тренд

$$w(z) = a_0 + a_1 z, \qquad (5)$$

то, рассчитав оценки коэффициентов $\tilde{a}_0$, $\tilde{a}_1$, из условия (4) получаем оценку искомого значения: $\tilde{z}_n = -a_0/a_1$. Аналогичным образом, полагая, что тренд описывается уравнением второго порядка

$$w(z) = a_0 + a_1 z + a_2 z^2, \qquad (6)$$

находим, что значению z_n при выполнении неравенства $d^2w/dz^2 > 0$ соответствует меньший из пары корней

$$z_{1,2} = \frac{-a_1 \pm \sqrt{a_1^2 - 4a_0a_2}}{2a_2}, \qquad (7)$$

а при справедливости соотношения $d^2w/dz^2 < 0$ – больший из корней (7) (рис. 1).

Если после нахождения оценки $\tilde{z}_n$ часть отброшенных ранее выборочных элементов либо все n_α элементов окажутся меньше $\tilde{z}_n$, их следует признать ложно принятыми за АД, пополнить ими исходную выборку, уточнить значения последних элементов ряда относительных частот $\{w_j\}$ и повторить процедуру оценивания z_n. При решении реальных задач влияние случайных погрешностей в значениях относительных частот $\{w_j\}$ обуславливает разброс оценок $\tilde{z}_n$, характеризуемый дисперсией $\sigma^2(\tilde{z}_n)$. Этот разброс следует учитывать при окончательном формировании величины значения z_n путем введения дополнительного смещения в уточненную оценку

$$\hat{z}_n = \tilde{z}_n + q\sigma(\tilde{z}_n),$$

где квантиль q обычно принимается близким к 2.

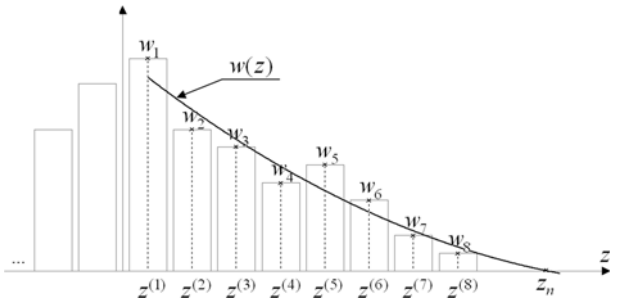

Рисунок 1 – Пример гистограммы интервального вариационного ряда, из которого формируются последовательности $\{z^{(j)}\}$, $\{w_j\}$

В принципе, для нахождения граничного значения z_n можно применять любые известные методы прогноза нестационарных временных рядов, в частности, экспоненциальное сглаживание, метод гармонических весов [7, 8]. Следует отметить, что последний, при справедливости предположения о монотонно убывающих «трендах» правой и левой сторон функции плотности вероятности основного распределения $\psi(z, \mu, \sigma_z^2)$, оказывается весьма адекватным для решения поставленной задачи.

ВЫЯВЛЕНИЕ АНОМАЛЬНЫХ ДАННЫХ МЕТОДОМ «ПРОГНОЗ – БУТСТРЕП»

При выявлении АД в выборке результатов наблюдений конечного объема n эффективным оказывается применение комбинированного подхода, состоящего в дополнении прогнозного метода, рассмотренного в предыдущем разделе, бутстреп-методологией тиражирования исходных данных [9]. Получение бутстреп-выборок $\{z_i\}_l^{\text{Б}}$, $l = \overline{1, L}$ позволяет избежать процедуры начального исключения из исходной выборки данных n_α «подозреваемых» на аномальность элементов. Блок-схема генерации бутстреп-выборок приведена в верхней части рис. 2.

Смысл процедуры бутстреп-генерации состоит в том, что из исходной выборки данных $\{z_i\}$, используемой в качестве генеральной совокупности, с помощью генератора ГПРЦЧ извлекаются методом выбора с возвращением отдельные элементы, из которых формируются псевдовыборки данных $\{z_i\}_l^{\text{Б}}$ (называемые также бутстреп-выборками), $l = \overline{1, L}$ объемом n элементов, статистически однородные исходной выборке $\{z_i\}$. Далее в каждой l-й бутстреп-выборке прогнозируется оценивание граничного значения z_n, из которых составляется выборка оценок $\{\tilde{z}_{nl}\}$, $l = \overline{1, L}$. Анализ этой выборки позволяет определить факт отсутствия либо наличия АД в исходной совокупности $\{z_i\}$ и при необходимости выделить эти АД. Суть процедуры установления факта наличия АД можно объяснить на примере обработки исходной выборки

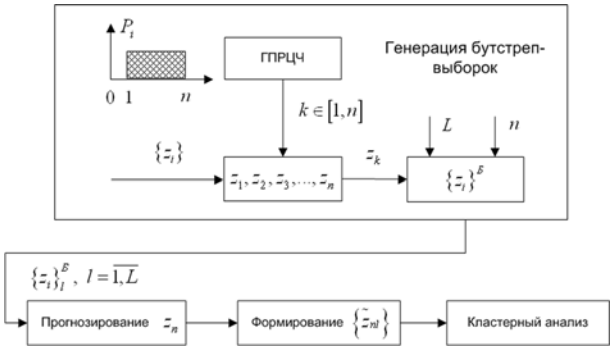

Рисунок 2 – Схема выявления АД применением комбинированной методики «прогноз – бутстреп»:

ГПРЦЧ – генератор псевдослучайных равномерно распределенных целых чисел

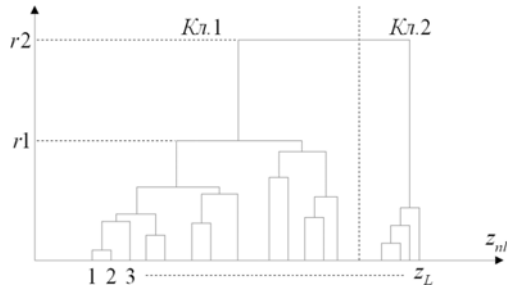

Рисунок 3 – Пример дендрограммы кластерного анализа данных, содержащих аномальности

$\{z_i\}$, содержащей один аномальный элемент. В этом случае при генерации бутстреп-данных общим объемом nL элементов вероятным будет появление в этом объеме L аномальных элементов, т. е. АД могут оказаться в идеальном варианте в каждой бутстреп-выборке либо (при попадании в бутстреп-выборки более одного аномального элемента) в меньшем количестве бутстреп-выборок. Прогноз граничных значений $\tilde{z}_n^{\alpha}$, полученный по выборке, содержащей АД, будет существенно отличен от прогнозных значений $\tilde{z}_n$, рассчитанных для выборок, составленных только из достоверных данных. Для обнаружения «аномальных» прогнозов $\tilde{z}_n^{\alpha}$ удобно применить кластерный анализ элементов выборки оценок $\{\tilde{z}_{nl}\}$, $l = \overline{1, L}$, допускающий визуализацию своих результатов в виде дендрограммы (рис. 3).

В случае наличия АД на дендрограмме будет легко визуально выделить два кластера: кластер Кл. 2 оценок «аномальных» прогнозов (группа из четырех оценок в правой части дендрограммы) и существенно более многочисленный кластер Кл. 1 оценок, найденных по достоверным данным. Расстояние $r2$, разделяющие эти кластеры, будет гораздо больше внутри кластерных расстояний. При отсутствии АД в исходной выборке результат кластерного анализа элементов выборки оценок прогнозов $\{\tilde{z}_{nl}\}$ покажет наличие в ней однородных данных, что выразится в формировании только одного кластера, объединяющего в себе все элементы $\tilde{z}_n$.

ВЫВОДЫ

Сложившаяся практика выявления аномальных данных опирается на так называемые смесевые модели засорения, позволяющие интерпретировать задачу выявления аномальных данных как задачу обнаружения элементов засоряющей совокупности. Однако ввиду невозможности достоверного оценивания вида и параметров засоряющего распределения и неадекватности реальным данным наиболее популярных моделей распределения основной совокупности наблюдений применение методов статистической классификации (в частности статистических методов распознавания и обнаружения) не позволяет получить корректное решение задачи выявления аномальных данных.

Предложенный подход позволяет получать информацию о виде распределения основной совокупности непосредственно из выборки и поэтому наиболее корректно формирует правило отбраковки аномальных данных.

ПЕРЕЧЕНЬ ССЫЛОК

1. *Жданюк Б. Ф.* Основы статистической обработки траекторных измерений / Жданюк Б. Ф. – М. : Сов. радио, 1978. – 384 с.
2. *Дубров А. М.* Многомерные статистические методы / Дубров А. М., Мхитарян В. С, Трошин Л. И. – М. : Финансы и статистика, 1998. – 352 с.
3. *Айвазян С. А.* Прикладная статистика: основы моделирования и первичная обработка данных : справочное изд. / Айвазян С. А., Енюков И. С., Мешалкин Л. Д. – М. : Финансы и статистика, 1983. – 471 с.
4. *Коваленко И. И.* Нетрадиционные методы статистического анализа данных / Коваленко И. И., Гожий А. П. – Николаев : Илион, 2006. – 116 с.
5. *Фомин А. Ф.* Отбраковка аномальных результатов измерений / Фомин А. Ф., Новоселов О. Н., Плющев А. В. – М. : Энергоатомиздат, 1985. – 200 с.
6. *Шурыгин А. М.* Прикладная статистика: робастность, оценивание, прогноз / Шурыгин А. М. – М : Финансы и статистика, 2000. – 224 с.
7. *Френкель А. А.* Прогнозирование производительности труда: методы и модели / Френкель А. А. – М. : Экономика, 1989. – 214 с.
8. *Лукашин Ю. П.* Адаптивные методы краткосрочного прогнозирования временных рядов / Лукашин Ю. П. – М. : Финансы и статистика, 2003. – 416 с.
9. *Эфрон Б.* Нетрадиционные методы многомерного статистического анализа / Эфрон Б. – М. : Финансы и статистика, 1988. – 263 с.

Надійшла 8.10.2008

Запропоновано непараметричний підхід до виявлення аномальних даних, що базується на прогнозному визначенні меж області довірчих значень вимірюваної величини. Наведено ряд методів, які дозволяють реалізувати викладений підхід на практиці.

The non-parametric method, based on the prognostic determination of reliable measured values limits is offered for detection of abnormal data. The set of methods allowing to solve the considered problem in practice is given.