УДК 519.25

А. В. ГОРОШКО, *кандидат технічних наук, доцент, доцент кафедри фізики і електротехніки Хмельницького національного університету, м. Хмельницький*

В. П. РОЙЗМАН, *доктор технічних наук, професор, старший науковий співробітник науково-дослідного відділу Національної академії Державної прикордонної служби України імені Богдана Хмельницького, м. Хмельницький*

УТОЧНЕННЯ МЕТОДУ ОБРОБКИ СТАТИСТИЧНИХ ДАНИХ

Проаналізовано методи статистичної обробки емпіричних даних в процесі вирішення проблем моніторингу стану прикордонної безпеки, надійності технічних засобів охорони державного кордону та ін. Запропонований метод уточненої обробки статистичних даних, подаючи закон розподілу у вигляді полімодальної суміші унімодальних законів. Метод дозволяє розкрити внутрішню структуру даних з урахуванням можливої полімодальності закону їх розподілу і поряд з обґрунтованим вибором кроку побудови гістограм дає правила роботи з такими даними. Подані результати практичного застосування методу.

Ключові слова: *статистична обробка, закон розподілу, полімодальність, суміш законів розподілу.*

Постановка проблеми у загальному вигляді. Обробка статистичних даних є важливою складовою процесу вирішення проблем моніторингу стану прикордонної безпеки, оцінки та прогнозування джерел та характеру загроз і викликів національним інтересам України на державному кордоні, розробки й обґрунтування методик оцінки ефективності оперативно-службової діяльності органів (підрозділів) охорони кордону прикордонного відомства, методик розрахунку потреб у силах і засобах, необхідних для ефективного виконання завдань за призначенням, обґрунтування методик оцінки ефективності функціонування управлінських структур органів (підрозділів) охорони державного кордону, надійності технічних засобів охорони державного кордону та ін.

Як видно з основних наукових робіт і дисертацій, захищених у спеціалізованих вчених радах Національної академії державної прикордонної служби України, найчастіше дослідники обробляють емпіричні дані, виходячи із параметричних статистичних гіпотез. Перевага застосування типових законів розподілу (нормального, логарифмічно нормального, експоненціального закону, закону Вейбулла, гамма-розподілу тощо) полягає в їх достатній вивченості та можливості отримання спроможних, незміщених і відносно високоефективних оцінок параметрів. В абсолютній більшості випадків пропонується вважати закон нормальним через достатньо загальні умови його появи – результат вимірювання (спостереження) складається під дією багатьох причин, причому кожна із них вносить лише малий внесок, сукупний загал визначається адитивно.

Аналіз останніх досліджень і публікацій, в яких започатковано вирішення даної проблеми та на які опирається автор. Часто гіпотеза про нормальність приймається після побудови гістограми, яка дозволяє "на око" оцінити нормальність розподілу та якісно оцінити деякі характеристики розподілу. Якщо гістограма має вигляд, наближений до "дзвону" Гауса, то її можна апроксимувати законом Гауса, але обов'язково частина емпіричних даних буде знаходитись поза кривої нормального закону. Такі значення беруться за випадкові викиди, після чого нормальність має бути підтверджена

перевіркою на нормальність для заданого рівня значимості за допомогою критеріїв нормальності (наприклад, критерію Колмогорова-Смирнова або W-критерію Шапіро-Уілка) [1, 2].

На відміну від прийнятого відомого способу автори пропонують здійснювати обробку даних, припускаючи, що спостереження, прийняті за "викиди", належать окремим модам, і закон розподілу насправді є не унімодальним, а полімодальним. Оскільки прийнявши закон розподілу даних полімодальним, не ясно, як його обробляти, автори пропонують такий метод уточненої обробки емпіричних даних.

Метою досліджень є розробка методу уточненої статистичної обробки емпіричних даних шляхом подання закону їх розподілу як суміші унімодальних законів розподілу, з подальшою її декомпозицією.

Виклад основного матеріалу дослідження.

Суть методу полягає у поданні й обробці емпіричної густини розподілу (ГР) у вигляді суперпозиції k функцій ГР f_i з вектором параметрів θ_i (компонент суміші), $i = 1,2,...,k$, $2 \le k < \infty$ у вигляді

$$f(x) = \sum_{i=1}^{k} \rho_i f_i(x, \theta_i), \qquad (1)$$

де $x \in \mathbb{R}$, ρ_i - апріорна імовірність (ваговий коефіцієнт) i-ї компоненти суміші, $\rho_i \in (0,1)$, $\sum_{i=1}^{k} \rho_i = 1$. У загальному випадку умова приналежності $\forall i$, $f_i(X, \theta_i)$ до одного параметричного сімейства не ставиться.

Нехай в результаті експерименту одержана вибірка значень $x = (x_1, x_2..., x_n)$. Для подальшої обробки результатів експерименту, перш за все, необхідно провести декомпозицію (розщеплення) суміші, тобто визначити невідомі параметри $\rho_1, \rho_2, ...\rho_{i-1}$, $\theta_1, \theta_2, ..., \theta_n$, наприклад, максимізуючи функцію максимальної правдоподібності [3].

$$W(\rho, \theta, x) = \prod_{j=1}^{n} \sum_{i=1}^{k} \rho_i f_i(x_j, \theta_i), \qquad (2)$$

прирівнюючи до нуля її частинні похідні за шуканими параметрами. Як правило, замість пошуку максимуму функції $W(\rho,\theta,x)$ простіше шукати максимум її логарифму

$$\ln W(\rho,\theta,x)=\sum_{j=1}^{n}\ln\left(\sum_{i=1}^{k}\rho_i f_i\left(x_j,\theta_i\right)\right), \qquad (3)$$

але навіть така постановка задачі без застосування спеціальних прийомів викликає значні труднощі. Тому для декомпозиції суміші (1) застосовують спеціальні методи: EM-алгоритм і його модифікації – SEM, CEM, MCEM, SAEM тощо; наближені методи, такі як метод фіксованих компонент з використанням методу найменших квадратів (МНК) та методу найменших модулів, а також Баєсовський класифікатор, описані, наприклад, в роботах [3–5].

Запропонований авторами метод декомпозиції сумішей базується на апроксимації функції ГР функцією типу (1) за допомогою МНК або інтерполяції на деякій точковій множині. У той же час відомо, що емпіричні дані вибірки $x=(x_1,x_2...,x_n)$ можуть бути подані лише варіаційним рядом, гістограмою або емпіричною функцією розподілу імовірностей

$$F_n(x)=\frac{1}{n}\sum_{j=1}^{n}1\left(x_j<x\right). \qquad (4)$$

Оскільки нормалізована гістограма (густина відносної частоти) при певних умовах є емпіричною ГР, побудованою для вибірки, апроксимація буде тим точнішою, чим краще побудована нормалізована гістограма буде наближатись до функції ГР імовірностей генеральної сукупності.

При побудові нормалізованих гістограм виникають певні труднощі. Оскільки вибір статистичної моделі розподілу визначається видом гістограми, який, у свою чергу, залежить від способу її побудови, і, особливо, від обраного кроку інтервалу значень, перед дослідниками природно постає питання, яким повинен бути крок розбиття h при побудові гістограм [6].

Зрозуміло, що зі зростанням кількості інтервалів гістограма не буде наближатись до ГР. Отже, існує деякий оптимальний крок $h_{>?B}$ побудови гістограм, при якому її апроксимація функцією (1) дасть оцінки параметрів ρ_i, θ_i, найближчі до їх справжніх значень.

Також є очевидним, що кількість компонент суміші k, визначених за гістограмою, залежить від кроку h. Для визначення h пропонується розглянути різні варіанти його значень, а отже і різні значення кількості компонент суміші k, і надалі вибрати оптимальне значення $h_{>?B}$, виходячи із деяких критеріїв якості. Одним зі способів може бути перебір всіх можливих значень h, k і оцінка отриманої моделі, максимізуючи деякі критерії, але такий спосіб є занадто ресурсозатратним.

Слід також пам'ятати, що з ростом k буде збільшуватись і правдоподібність моделі (3), оскільки більш гнучка модель може краще пояснити досліджувані дані, тому цю задачу неможливо розв'язати, просто шукаючи k із умови максимуму правдоподібності і включивши його до шуканих параметрів.

Для вибору оптимального кроку $h_{опт}$ авторами запропонований наступний ітераційний алгоритм. Крок має бути мінімальним, але не менше за точність вимірювання параметра ε. Оскільки справжня кількість мод є невідомою, пропонується вибирати початкову кількість компонент k в (1) наперед більшою, наприклад такою, що дорівнює кількості локальних максимумів функції $f(x)$.

Далі запропонованим раніше методом необхідно визначити невідомі параметри $\rho_1, \rho_2, ...\rho_{i-1}, \theta_1, \theta_2, ..., \theta_n$. Якщо в результаті розрахунків один або декілька вагових коефіцієнтів ρ_i виявляться менше деякої наперед заданої порогової величини β, то відповідними членами у лінійній комбінації (1) можна знехтувати.

Далі крок гістограми можна збільшувати до тих пір, поки кількість вершин (локальних максимумів) не стане дорівнювати кількості членів k в лінійній комбінації (1) після відкидання її малих членів. Знову застосовуючи той же метод розв'язку, але вже для меншої кількості невідомих, можна визначити їх уточнене значення і відкинути малі члени. Такий процес слід продовжувати до тих пір, поки всі ρ_i не ста-

нуть порівнювані з вибраною точністю β. Отриманий при цьому крок може бути взятий за оптимальний $h_{опт}$. Фізично цей процес означає, що підвибірки з малим ρ_i вносять вельми незначний внесок у загальну вибірку, і тому їх можна об'єднати з однією із підвибірок виробів з близькими величинами досліджуваного параметра.

Результати практичного застосування

Розглянемо гістограму відносних частот, одержану в результаті обробки деякої вибірки статистичних даних у кількості $n = 583$, поданих у вигляді варіаційного ряду. Зображення нормалізованої гістограми, побудованої з кроком $h = 2{,}5$, наведене на рис. 1. Прийнявши гіпотезу про нормальність, одержано функцію нормального розподілу з параметрами $\mu = 29{,}7$, $\sigma = 9{,}6$, графік якої поданий на рис. 1.

Застосуємо описаний вище метод уточнення статистичної обробки шляхом подання емпіричного закону розподілу у вигляді суміші трьох нормальних законів розподілу. Для цього використаємо всі наявні дані варіаційного ряду. У результаті декомпозиції суміші одержимо результуючу ГР типу (1) з параметрами $\rho_1 = 0{,}025$, $\mu_1 = 13$, $\sigma_1 = 1{,}8$, $\rho_2 = 0{,}94$, $\mu_2 = 29{,}7$, $\sigma_2 = 9{,}6$, $\rho_3 = 0{,}035$, $\mu_3 = 33$, $\sigma_3 = 1{,}5$, графік якої зображений на рис. 2.

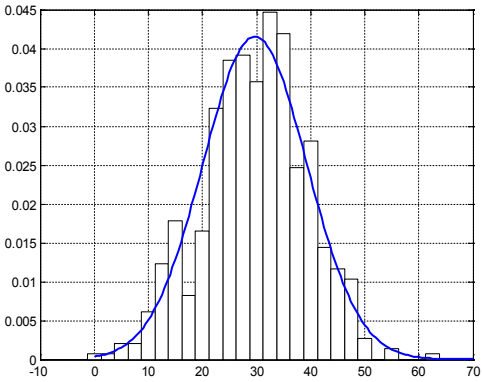

Рис. 1. Гістограма відносних частот

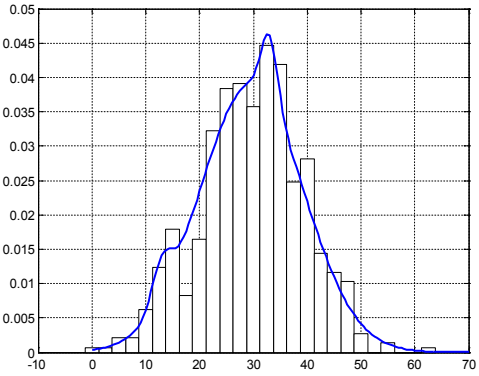

Рис. 2. Графік ГР суміші трьох нормальних законів

Якщо ж дослідника задовольняє менша точність, емпіричну ГР можна прийняти двомодальною і провести декомпозицію суміші для двох нормальних законів. На рис. 3 і 4 подані результати розщеплення суміші розподілу для двох різних апріорних імовірностей ρ основної моди.

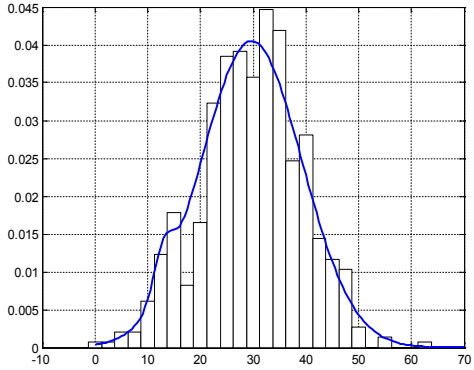

Рис. 3. Графік ГР суміші двох нормальних законів з параметрами $\rho_1 = 0{,}025$, $\mu_1 = 13$, $\sigma_1 = 1{,}8$, $\rho_2 = 0{,}975$, $\mu_2 = 29{,}7$, $\sigma_2 = 9{,}6$

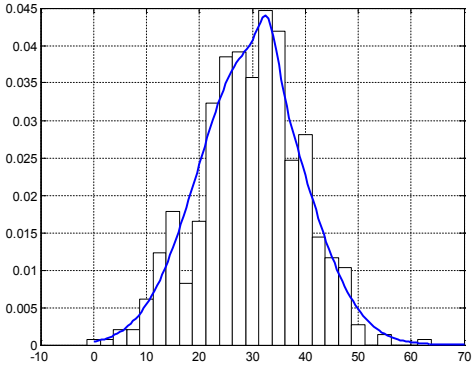

Рис. 4. Графік ГР суміші двох нормальних законів з параметрами $\rho_1 = 0,03$, $\mu_1 = 33$, $\sigma_1 = 1,8$, $\rho_2 = 0,97$, $\mu_2 = 29,7$, $\sigma_2 = 9,9$

Висновки. У роботі показаний метод уточненої обробки емпіричних статистичних даних, подаючи закон розподілу у вигляді полімодальної суміші одномодальних законів. Метод дозволяє розкрити внутрішню структуру даних із врахуванням можливої полімодальності закону їх розподілу і поряд з обґрунтованим вибором кроку побудови гістограм дає правила роботи з такими даними. Подані результати практичного застосування методу.

Перспективою подальших розвідок у даному напрямку є розробка методів призначення довірчих інтервалів для полімодальних законів розподілу та їх практичне застосування для визначення допустимих значень даних спостережень.

Список використаної літератури

1. Вентцель Е.С. Теория вероятностей / Е.С. Вентцель – М.: Наука, 1969. – 576 с.

2. Орлов А. И. Прикладная статистика. Учебник. / А. И. Орлов. – М. : Издательство "Экзамен", 2004. – 656 с.

3. Прикладная статистика: Классификация и снижение размерности: Справ. изд. / С. А. Айвазян, В. М. Бухштабер, И. С. Енюков,

Л. Д. Мешалкин ; под ред. С. А. Айвазяна. – М.: Финансы и статистика, 1989. – 607 с. : ил.

4. Королев В. Ю. ЕМ-алгоритм, его модификации и их применение к задаче разделения смесей вероятностных распределений. Теоретический обзор / В. Ю. Королев. – М. : Изд-во ИПИ РАН, 2007.

5. S. F. Nielsen. The stochastic EM algorithm: estimation and asymptotic results. – Bernoulli, 2000, vol. 6, No. 3, p. 457–489.

6. Горошко А. В. Представление и обработка статистических данных, не подчиняющихся унимодальным законам распределения / А. В. Горошко, В. П. Ройзман // Машиностроение и инженерное образование, 2013. – № 3. С. 56–77.

Стаття надійшла до редакції 26.02.2014.

Горошко А. В., Ройзман В. П. **Уточнение метода обработки статистических данных**

Проанализированы методы статистической обработки эмпирических данных в процессе решения проблем мониторинга состояния пограничной безопасности, надежности технических средств охраны государственной границы и др. Предложен метод уточненной обработки статистических данных, представляя закон распределения в виде полимодальной смеси одномодальных законов. Метод позволяет раскрыть внутреннюю структуру данных с учетом возможной полимодальности закона их распределения и наряду с обоснованным выбором шага построения гистограмм дает правила работы с такими данными. Представлены результаты практического применения метода.

Ключевые слова: статистическая обработка, закон распределения, полимодальность, смесь законов распределения.

Goroshko A.V., Royzman V. P. **Refined methods of statistical data processing**

Handling statistical data is an essential component to solving the problems of monitoring the state of border security, assessment and prediction of the sources and nature of the threats and challenges to national interests at the state border of Ukraine, development and study

of methods for evaluating the effectiveness of the operational activities of the border agency, techniques calculation needs men and equipment necessary for the effective performance of assigned tasks, study methodologies to assess the efficiency of the management structures of the (sub) state border protection , reliability, technical means of protection of the state border, and others.

In contrast to the known method adopted by the authors propose to treat the data, suggesting that the observation made by "emissions" are those of the individual modes and distribution law really is not unimodal and polymodal.

The method is presented and processed empirical density distribution as a superposition of features unimodal density distribution (mixing) with the a priori probabilities (weights).

The authors propose a method based on decomposition of mixtures approximation of probability density by the method of least squares interpolation or some dotted-set.

The paper presents a method of processing refined empirical statistical data representing the distribution law as polymodal mixture unimodal laws. The method can reveal the internal structure of the data, taking into account possible polimodal law and their distribution, along with a reasonable selection step histogram gives the rules for dealing with such data.

In subsequent publications, the authors will describe the method for assigning confidence intervals for polymodal distribution laws and their practical application to determine the acceptable values of observational data.

Keywords: *statistical processing, distribution law, polymodal, the mixture distribution laws.*