УДК 519.25

І.О. ДУДКО, аспірант, КПУ, Запоріжжя,
В.Є. БАХРУШИН, д-р физ.-мат. наук, проф., КПУ, Запоріжжя

ВИКОРИСТАННЯ МЕТОДУ *k*-СЕРЕДНІХ ДЛЯ ІДЕНТИФІКАЦІЇ МОДЕЛЕЙ НЕОДНОРІДНИХ РОЗПОДІЛІВ ВИПАДКОВИХ ВЕЛИЧИН

У статті розглянуто проблему побудови моделей неоднорідного розподілу випадкових величин. Запропоновано процедуру, що передбачає попереднє оцінювання кількості та параметрів складових та їх подальше уточнення методами мінімізації розрахункових значень критерію Колмогорова–Смирнова або критерію типу омега-квадрат. Наведено результати аналізу емпіричних даних про вибори Президента України у 2010 р. Іл.: 3. Табл.: 1. Бібліогр.: 11 назв.

Ключові слова: неоднорідний розподіл, ідентифікація, метод *k*-середніх, критерій Колмогорова-Смирнова.

Постановка проблеми та аналіз літератури. Генеральна сукупність, утворена об'єднанням двох чи більше однорідних вибірок є неоднорідною і зазвичай не може бути описана будь-яким однорідним законом розподілу.

Нехай $p_k \geq 0$, $\sum_k p_k = 1$, $F_k(x)$ − деякі функції розподілу. Функція розподілу неоднорідної випадкової величини є сумішшю функцій $F_k(x)$ з вагами p_k [1]

$$F(x) = \sum_k p_k F_k(x). \qquad (1)$$

Постає питання розщеплення суміші розподілів на складові. Розв'язати цю задачу − означає за вибіркою класифікованих спостережень

$$X_1, X_2, ..., X_n, \qquad (2)$$

отриманих з неоднорідної генеральної сукупності типу (1), визначити кількість k компонентів суміші, вагові коефіцієнти $p_1, p_2, ..., p_k$ та параметри функції $F_k(x)$.

Задача ідентифікації неоднорідних розподілів випадкових величин часто виникає в прикладних задачах, пов'язаних з поділом або розщепленням сукупностей даних [2], наприклад, в задачах класифікації або розпізнавання образів. В статті О.К. Ісаєнка та В.Ю. Урбаха [3] розглядаються питання, пов'язані з розділенням сумішей розподілів ймовірностей на їх складові класичними методами та методами

кластерного аналізу. Різноманітні засоби вирішення задачі класифікації об'єктів наводяться в роботі С.А. Айвазяна [4]. Питання пов'язані зі статистичними критеріями перевірки адекватності моделей вирішуються в роботах О.І. Орлова [5, 6], Б.Ю. Лемешка [7], К.В. Воронцова [8] та інших науковців.

Мета статті. Розробка алгоритму ідентифікації моделей неоднорідного розподілу випадкових величин з використанням структурно-параметричного підходу.

Побудова математичної моделі. Ідентифікація закону розподілу випадкової величини полягає у виборі параметричної моделі закону розподілу ймовірностей, що найточніше відповідає результатам експериментальних спостережень [7].

Структурно-параметричний підхід до ідентифікації закону розподілу складається з двох етапів [9]:

1. На основі аналізу вибіркових даних будують параметричну модель досліджуваного розподілу і роблять оцінки її параметрів.
2. За допомогою мінімізації розрахункових значень критеріїв Колмогорова–Смирнова або омега-квадрат уточнюють параметри моделі.
3. Перевіряють адекватність отриманої моделі за допомогою статистичних критеріїв.

Перший етап можна реалізувати за допомогою методу k-середніх кластерного аналізу, який було запропоновано Д. Хартиганом і М. Вонгом у 1978 році [5].

Алгоритм k-середніх передбачає виконання таких основних етапів [5, 10].

Етап 1. Формують k початкових центрів кластерів $y_1(1)$, $y_2(1), \ldots, y_k(1)$.

Етап 2. На n-му кроці ітерації множину об'єктів x поділяють на k кластерів за таким правилом:

$$x \in S_i(n), \text{ якщо } \left\|x - y_j(n)\right\| < \left\|x - y_i(n)\right\|$$

для всіх $i = 1, 2, \ldots, k$, $i \neq j$, де $S_i(n)$ – множина об'єктів, що належать до кластера з центром $y_i(n)$, функцією відстані є евклідова метрика.

Етап 3. Визначають нові центри кластерів $y_j(n+1)$, $j = 1, 2, \ldots, k$, з умови, що:

$$J_j = \arg\min_{x \in S_j(n)} \left\|x - y_j(n+1)\right\|^2, \; j = 1, 2, \ldots, k,$$

де J_j – показник якості мінімізації.

Центр $y_i(n+1)$ є вибірковим середнім на множині $S_j(n+1)$, тому нові центри кластерів визначають як

$$y_j(n+1) = \frac{1}{N_j} \sum_{x \in S_j(n)} x, \quad j = 1,\ 2,\ \ldots,\ k,$$

де N_j – кількість об'єктів вибірки, що належать до множини $S_j(n)$.

Етап 4. Умовою збіжності алгоритму є виконання рівності $y_j(n+1) = y_j(n)$ при $j = 1, 2, \ldots, k$. Якщо ж ця умова не виконується, то необхідно повернутися до етапу 2.

Однією з головних проблем кластерного аналізу методом k-середніх є вибір початкової кількості кластерів. Для її вирішення можна використовувати ієрархічну класифікацію.

Після визначення кількості кластерів та їх статистичних характеристик уточнюють параметри моделі розподілу (1) шляхом мінімізації розрахункових значень критерію Колмогорова–Смирнова або омега-квадрат [6, 11].

На третьому етапі необхідно перевірити відповідність даних вибірки знайденій моделі розподілу. Для цього доцільно використовувати ті самі критерії, значення яких мінімізували на попередньому етапі.

Результати аналізу емпіричних даних. Для демонстрації роботи процедури ідентифікації моделей неоднорідних розподілів випадкових величин були використані результати голосування за кандидата В.Ф. Януковича на чергових виборах президента України від 17.01.2010 р. [12].

Вибірка складається з результатів голосування по 226 округам. Аналіз даних виконували за допомогою статистичних пакетів SPSS, Statistica [13, 14] та MS Excel. Були знайдені статистичні значення вибірки (табл.), побудовані гістограма та функція розподілу (рис. 1).

Таблиця

Статистичні показники результатів голосування

Обсяг вибірки	226	Стандартне відхилення	23,37
Розмах	81,08	Дисперсія	545,93
Мінімум	3,09	Асиметрія	0,406
Максимум	84,17	Ексцес	–1,14
Середнє значення	36,10	Коефіцієнт варіації	0,65

Виходячи з того, що коефіцієнт варіації більше ніж 0,5, а також з аналізу P-P графіків та гістограми, можна зробити висновок про неоднорідність досліджуваного розподілу.

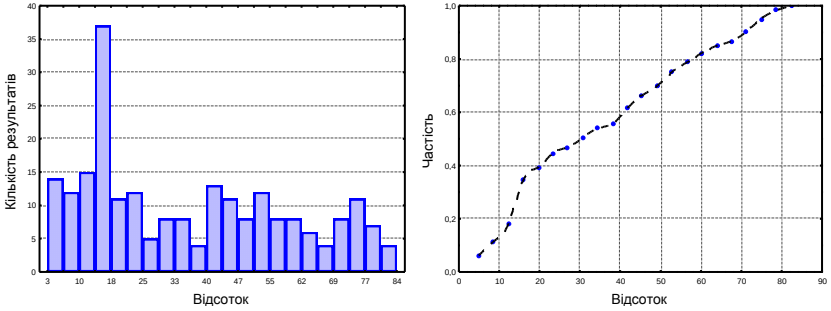

Рис. 1. Гістограма та емпірична функція розподілу результатів голосування

Перед тим, як перейти до методу k-середніх, необхідно визначити кількість однорідних компонент вибірки, тобто кластерів. Для цього необхідно побудувати дендрограму на основі вибіркових значень (рис. 2).

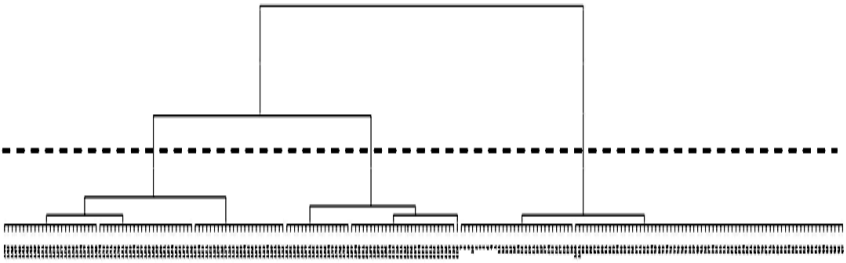

Рис. 2. Дендрограма розподілу результатів голосування

Поділивши денрограму навпіл, отримаємо перетин з трьома групами значень. Отже кількість кластерів $k = 3$.

Далі здійснюємо процедуру кластеризації методом k-середніх. В результаті отримаємо 3 кластери, що складаються зі 107, 58 та 61 елементів. На основі аналізу P-P графіків та статистики Колмогорова-Смирнова, можна стверджувати, що найбільш придатним для опису кожної групи значень є нормальний розподіл з параметрами $N(14{,}90;\ 6{,}30)$, $N(43{,}79;\ 7{,}80)$, $N(70{,}31;\ 7{,}41)$. Якщо мінімізувати суму квадратів різниць значень емпіричних та теоретичних функцій розподілу для кожного кластера, отримаємо такі уточнення: $N(14{,}70;\ 6{,}23)$,

$N(47{,}62; 13{,}41)$, $N(75{,}09; 4{,}42)$. Вагові коефіцієнти дорівнюють: $p_1 = 0{,}45$, $p_2 = 0{,}42$, $p_3 = 0{,}13$. Таким чином, отримана модель розподілу результатів голосування є такою:

$$F_M = 0{,}45N(14{,}70; 6{,}23) + 0{,}42N(47{,}62; 13{,}41) + 0{,}13N(75{,}09; 4{,}42).$$

Розрахункове значення критерію Колмогорова−Смирнова дорівнює 0,41 і є меншим ніж критичне 0,895. Отже, можна вважати, що емпірична й теоретична функції розподілу відповідають одна одній на рівні значимості 0,05. Добра відповідність між емпіричною та теоретичною функціями розподілу також ілюструється рис. 3.

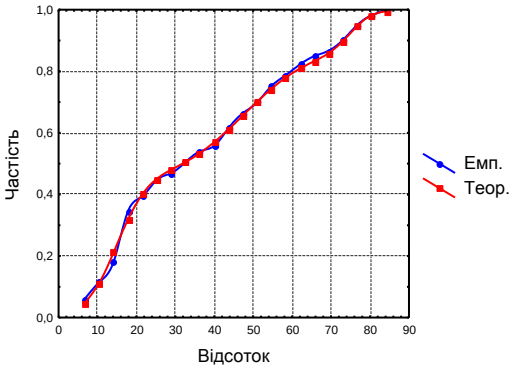

Рис. 3. Порівняння емпіричної та теоретичної функції розподілу результатів голосування

Висновки. Розроблено алгоритм ідентифікації моделей неоднорідного розподілу випадкових величин з використанням структурно-параметричного підходу. На першому етапі пропонується визначати кількість компонент та початкове наближення їх статистичних характеристик за допомогою методів кластерного аналізу. На другому етапі ідентифікації запропоновано уточнювати параметри моделі розподілу шляхом мінімізації розрахункових значень непараметричних критеріїв згоди (Колмогорова-Смирнова, омега-квадрат), а на третьому – перевіряти адекватність отриманої моделі за допомогою відповідних критеріїв. Результати обчислювальних експериментів підтверджують можливість використання запропонованого алгоритму.

Список літератури: 1. *Ивченко Г.И.* Введение в математическую статистику: Учебник / *Г.И. Ивченко, Ю.И. Медведев*. − М.: Издательство ЛКИ, 2010. − 600 с. **2.** *Королев В.Ю.* Математические основы теории риска / *В.Ю. Королев, В.Е. Бенинг, С.Я. Шоргин*. − М.: Физматлит, 2007. − 544 с. **3.** *Исаенко О.К.* Разделение смесей распределений вероятностей

на их составляющие / *О.К. Исаенко, В.Ю. Урбах* // Итоги науки и техники. Сев. ТВ. МС. ТК ВИНИТИ, 1976. − Т. 13. − С. 37 − 58. **4.** *Айвазян С.А.* Прикладная статистика: Классификация и снижение размерности. Справ. изд. / *С.А. Айвазян, В.М. Бухштабер, Е.С. Енюков*; Под ред. *С.А. Айвазяна*. − М.: Финансы и статистика, 1989. − 607 с. **5.** *Орлов А.И.* Математика случая: Вероятность и статистика – основные факты: Учебное пособие / *А.И. Орлов.* – М.: МЗ-Пресс, 2004. – 110 с. **6.** *Орлов А.И.* Прикладная статистика / *А.И. Орлов* − М.: Экзамен, 2006. − 671 с. **7.** *Лемешко Б.Ю.* О задаче идентификации закона распределения случайной составляющей погрешности измерений / *Б.Ю. Лемешко* // Метрология. − 2004. − № 7. – С. 8 − 18. **8.** *Воронцов К. В.* Качество восстановления зависимостей по эмпирическим данным / *К. В. Воронцов* // Математические методы распознавания образов: 7-ая Всерос. конф. Тезисы докл. – Пущино, 1995. – С. 24 – 26. 9**.** *Бахрушин В.Е.* Проблемы идентификации моделей распределения случайных величин с применением современного программного обеспечения / *В.Є. Бахрушин* // Успехи современного естествознания. – 2011. – № 11. – С. 50 – 54. **10.** *Ту Дж.* Принципы распознавания образов / *Дж. Ту, Р. Гонсалес*. − М. : Мир, 1978. – 414 с. **11.** *Бахрушин В.Є.* Методи аналізу даних / *В.Є. Бахрушин*. – Запоріжжя: КПУ, 2011. – 268 с. **12.** Основні статистичні відомості виборчого процесу 2010 року з чергових виборів Президента України // Вісник Центральної виборчої комісії. – 2010. – № 1. – С. 12 − 15. **13.** *Бююль А.* SPSS: искусство обработки информации. Анализ статистических данных и восстановление скрытых закономерностей / *Ахим Бююль, Петер Цёфель*. − СПб.: ООО "ДиаСофтЮП", 2005. − 608 с. **14.** *Халафян А.А.* STATISTICA 6. Статистический анализ данных. 3-е изд. Учебник / *А.А. Халафян*. − М.: ООО "Бином-Пресс", 2007. − 512 с.

УДК 519.25

Использование метода *k*-средних для идентификации моделей неоднородных распределений случайных величин / Дудко И.А., Бахрушин В.Е. // Вестник НТУ "ХПИ". Серия: Информатика и моделирование. − Харьков: НТУ "ХПИ". − 2012. − №. 62 (968). – С. 64 − 69.

В статье рассмотрена проблема идентификации моделей неоднородных распределений случайных величин. Предложена процедура, которая предусматривает предварительное оценивание количества и параметров составляющих и их дальнейшее уточнение методами минимизации расчетных значений критерия Колмогорова−Смирнова или критерия типа омега-квадрат. Приведены результаты анализа эмпирических данных о выборах Президента Украины в 2010 г. Ил.: 3. Табл.: 1. Библиогр.: 14 назв.

Ключевые слова: неоднородное распределение, идентификация, метод *k*-средних, критерий Колмогорова−Смирнова.

UDC 519.25

Using of the *k*-means method for identification of inhomogeneous distributions models of random variables / Dudko I.O., Bakhrushin V.E. / Herald of the National Technical University "KhPI". Subject issue: Information Science and Modeling. – Kharkov: NTU "KhPI". – 2012. – №. 62 (968). – P. 64 – 69.

In the article the problem of identification of inhomogeneous distribution models of random variables is considered. The procedure, which provides a preliminary estimation of the quantity and parameters of components and their further improvement with the help of methods of minimization of the Kolmogorov−Smirnov test or the omega-squared test calculated values, is proposed. The results of analysis of empirical data about the elections of the President of Ukraine in 2010 year. Figs.: 3. Tabl.: 1. Refs: 14 titles.

Keywords: inhomogeneous distribution, identification, *k*-means method, the Kolmogorov−Smirnov test.

Надійшла до редакції 02.08.2012