враховувати численні фактори та адаптуватися до змінних вподобань користувачів. Це підвищує задоволеність користувачів та сприяє більш ефективному використанню платформи.

Результати дослідження підтверджують, що розвиток систем рекомендацій у сфері кінематографії рухається в бік використання інтелектуального аналізу даних, що дає змогу підвищити їх ефективність і точність. Перспективи подальших досліджень включають розробку нових алгоритмів, які зможуть ще краще враховувати індивідуальні потреби та контекстуальні фактори, а також дослідження етичних аспектів використання таких систем.

Список використаних джерел

1. Ankit J. Role of a Movie Recommender System in the Streaming Industry. 2022. URL: https://www.muvi.com/blogs/movie-recommender-system/

2. Щербань В. С. Інформаційна технологія відбору відеоматеріалів. Програмні засоби на основі колаборативної фільтрації. URL: https://inmad.vntu.edu.ua/portal/static/D2FE0A0F-A676-4969-A811-F7F6946D54A4.pdf

3. Prabhat. Creating a Hybrid Neural Network for Movie Recommendations using TensorFlow Recommenders. 2023. URL: https://prabhatm27.medium.com/creating-a-hybrid-neural-network-for-movie-recommendations-using-tensorflow-recommenders-fdad13aa4979

4. Як працює система рекомендацій Netflix: *Netflix.* URL: https://help.netflix.com/uk/node/100639

УДК 004.852

Семенюк А. М., *здобувач 3 курсу спеціальності 122 Комп'ютерні науки,*
Хмелівський Ю. С., *асистент кафедри інформаційних технологій*

КЛАСТЕРНИЙ АНАЛІЗ СТАТИСТИЧНИХ МЕДИЧНИХ ДАНИХ НА МОВІ R

Донецький національний університет імені Василя Стуса, м. Вінниця

Вступ. Статистичний аналіз даних засобами мови R дає змогу розв'язувати основні статистичні задачі, візуалізувати дані, виконувати аналіз даних та прогнозування результату.

Засоби для візуалізації результатів обчислень дають змогу створювати різного виду графіки, що сприяють легкому сприйняттю інформації.

Значні можливості мови R для здійснення статистичних аналізів пов'язані з наявністю засобів лінійної і нелінійної регресії, класичних статистичних тестів, часових рядів (серій), кластерних обчислень і багато іншого [1].

Для успішного проведення статистичних досліджень важливо мати методологію, яка дає змогу проводити збір даних, що містять інформацію про вибірку об'єктів і потім упорядковує їх у відносно однорідні групи. Мова R має такий інструмент – це кластерний аналіз.

Перелік напрямів та галузей, де може бути використаний кластерний аналіз, дуже великий. Насамперед це пов'язано з головною метою такого дослідження – знаходження груп схожих об'єктів у вибірці.

Статистичний аналіз даних на мові R. У цій роботі розглядаються можливості цієї методології для обробки статистичних даних у медицині, зокрема – аналіз діяльності медико-соціальних експертних комісій України з питань медико-соціальної реабілітації осіб з інвалідністю [2].

Для забезпечення планової роботи медико-соціальної експертизи та реабілітації осіб з інвалідністю статистичні методи дають змогу виробити обґрунтовані, правильні організаційні та методологічні рішення, що пов'язане з ретельним аналізом отриманої інформації.

Статистичне дослідження як процес можна розбити на п'ять основних етапів:

1. Визначення мети та завдань дослідження.
2. Збір статистичного матеріалу.
3. Попередня обробка даних.
4. Розрахунок та інтерпретація узагальнюючих статистичних показників.
5. Моделювання та прогнозування.

Перші три етапи не вимагають використання математичного апарату. П'ятий ґрунтується на даних обробки, отриманих на попередньому етапі. Використання функцій мови R дає змогу виконувати розрахунки показників та їх різносторонню інтерпретацію з формуванням вихідних форм [3, 4]. Як приклад кластерного аналізу наведено результуючу таблицю для формування рангів показників реабілітації (рис. 1) в розрізі областей.

Як уже зазначалось, наявність великих об'ємів інформації потребує особливого підходу до її обробки. По-перше, необхідно з'ясувати, як це виконати. Найефективніше – виконати сортування на групи для можливості попереднього аналізу. Групування за методом кластеризації – це метод, коли в самому алгоритмі відбору закладено пошук ознак, які об'єднують дані за визначеним критерієм. Тобто алгоритм із первинного потоку інформації виявляє параметри з подібними характеристиками.

Внаслідок виконання цієї процедури ми отримаємо масив даних розбитий на окремі кластери (підмножини). Елементи у кластерах мають схожі ознаки, водночас мають суттєву відмінність від елементів іншого кластера – підмножини між собою не перетинаються.

Кластери можна утворювати, ґрунтуючись на відстані між ними, на щільності ділянок у просторі даних, інтервалах або на конкретних статистичних розподілах [5]. Формування їх критеріїв – це ітераційний процес, він залежить від мети використання результатів і від самого вхідного набору інформації. Цей потік може мати будь-який розподіл даних:

- ✓ порядковий;
- ✓ інтервальний;
- ✓ категорійний.

Допускається також об'єднання типів, але в такому випадку ускладнюється аналіз (кластеризація).

Рангові місця служби МСЕ областей України
за показниками повної реабілітації та реабілітації інвалідів ІІІ групи
за 2022 рік (на 100 переоглянутих) [1]

Рангові місця	Адміністративні території	Показник повної реабілітації	Рангові місця	Адміністративні території	Не визнані інвалідами серед переоглянутих інвалідів ІІІ групи
1	Дніпропетровська	3,27	1	м. Київ	5,38
2	м. Київ	2,77	2	Дніпропетровська	4,43
3	Запорізька	2,70	3	Запорізька	3,78
4	Вінницька	2,47	4	Вінницька	3,03
4	Черкаська	2,47	4	Черкаська	3,03
6	Миколаївська	2,21	6	Миколаївська	2,89
7	Кіровоградська	1,94	7	Київська	2,71
8	Київська	1,71	8	Кіровоградська	2,69
9	Херсонська	1,41	9	Херсонська	2,12
10	Сумська	1,34	10	Сумська	1,87
11	Тернопільська	1,26	11	Одеська	1,70
12	Луганська	1,23	12	Тернопільська	1,65
13	Полтавська	1,06	13	Полтавська	1,55
14	Волинська	0,90	14	Луганська	1,45
15	Одеська	0,88	15	Волинська	1,29
16	Закарпатська	0,87	16	Донецька	1,17
17	Донецька	0,85	17	Закарпатська	1,16
18	Хмельницька	0,83	18	Хмельницька	1,08
19	Чернівецька	0,76	19	Чернівецька	1,01
20	Львівська	0,69	20	Львівська	0,84
21	Харківська	0,58	21	Харківська	0,83
22	Чернігівська	0,55	22	Чернігівська	0,76
23	Житомирська	0,51	23	Житомирська	0,74
24	Рівненська	0,49	24	Рівненська	0,61
25	Івано-Франківська	0,14	25	Івано-Франківська	0,19
	В Україні, 2022 р.	1,62		В Україні, 2022 р.	2,27

Рис. 1. Зразок таблиці вихідних даних

Математично кластеризація описується виразом:

$$S(c) = \sum_{i=1}^{N} s(i, c_i) + \sum_{k=1}^{N} \delta(c_k), \quad (1)$$

де c_k – множина вхідних елементів; $s(i, c_i)$ – множина подібності; $\delta(c_k)$ – символ обмежувач; k – критерій відбору (середина діапазону ранжування).

У подібному алгоритмі розподіл на кластери можна представити як завдання дискретної максимізації з деякими обмеженнями.

Вибір мови R пов'язаний з її орієнтацією на розв'язок і аналіз статистичних задач. Під час її роботи можливо виконувати одразу багато інструкцій, що записані в окремому файлі (скрипті) і збираються у пакети (packages) користувача.

Середовище розробки та наявні базові пакети дають змогу провести кластерний аналіз за допомогою алгоритму *k*-відбору, а можливості графіки покращують сприйняття результатів. Результатом задачі кластеризації з використанням цього алгоритму є векторні набори даних. Робота з програмами на мові R інтуїтивно проста та не потребує спеціальних навичок, а пакет доступний на безкоштовній основі, що сприяє його використанню.

Висновки. Медико-соціальну експертизу в Україні у 25 областях здійснювали 362 МСЕК. Об'єм звітних даних, які вони підготовляють, значний. Тому одним з основних завдань МСЕК, особливо на сучасному етапі реформування охорони здоров'я в Україні, є регулярне проведення організаційно-методичної роботи, яка неможлива без аналітичної інформації. Підготовка таких даних, їх інтерпретація та групування за потрібними ознаками може бути виконана з використанням програми кластерного аналізу на мові R навіть особами з базовою комп'ютерною підготовкою.

Список використаних джерел

1. Семенюк А. М., Хмелівський Ю. С. Статистичний аналіз медичних даних на мові R. *Прикладні аспекти сучасних міждисциплінарних досліджень*: матеріали II Міжнародної науково-практичної конференції (м. Вінниця, 24 листопада 2023 р.). Вінниця: ДонНУ імені Василя Стуса. 2023. 282 с.

2. Основні показники медико-соціальної реабілітації осіб з інвалідністю в Україні за 2022 рік. Аналітико-інформаційний довідник / В. І. Шевчук, Р. Я. Перепелична, Л. О. Сторожук, І. В. Куриленко, Л. Г. Семененко, М. В. Семенюк, А. М. Семенюк. Вінниця: ФОП Данилюк В. Г. 2023. 119 с.

3. Методи програмування в R: вебсайт. URL: https://tvimc.jimdofree.com (дата звернення: 18.05.2024).

4. Селезньов О. Мова R для користувачів Excel. 2022. URL: https://selesnow.github.io/r4excel_users/index.html (дата звернення: 18.05.2024).

5. Junkui L., Yuanzhen W., Xinping L. LB HUST: A symmetrical boundary distance for clustering time series. *9th International Conference on Information Technology (ICIT'06)*. 2006. С. 203–208.