Застосування Кластерного Аналізу у Дослідженні Особливостей Зоряного Неба

Дмитро Кобзев
інститут Комп'ютерного моделювання, прикладної фізики та математики
Національний технічний університет «Харківський політехнічний інститут»
Харків, Україна
dmytro.o.kobzev@infiz.khpi.edu.ua

Валерій Успенський
кафедра Комп'ютерного моделювання процесів та систем
Національний технічний університет «Харківський політехнічний інститут»
Харків, Україна
valerii.uspenskyi@khpi.edu.ua

Application of Cluster Analysis in the Study of the Starry Sky's Peculiarities

Dmytro Kobzev
Institute of Computer Modeling, Applied Physics and Mathematics
National Technical University "Kharkiv Polytechnic Institute"
Kharkiv, Ukraine
dmytro.o.kobzev@infiz.khpi.edu.ua

Valerii Uspenskyi
Department of Computer Modeling of Processes and Systems
National Technical University "Kharkiv Polytechnic Institute"
Kharkiv, Ukraine
valerii.uspenskyi@khpi.edu.ua

***Анотація*—Аналізуються застосування відомих методів кластеризації стосовно об'єктів спостереження зоряного неба. Для досліджень використані мова програмування Python і фреймворк Keras.**

***Abstract*—The application of known clustering methods for the starry sky objects observation is analyzed. The Python programming language and the Keras framework were used for research.**

Ключові слова— об'єкти спостереження зоряного неба, методи кластеризації, нейронна мережа Кохонена

Keywords—star sky observation objects; clustering methods; Kohonen neural network

I. Вступ

Фіксація параметрів об'єктів спостереження зоряного неба виконується цілодобово міжнародною мережею телескопів, що охоплює територію усіх континентів Земної кулі та острови у її морях і океанах. Переважна більшість з них встановлена у обсерваторіях, що входять до Міжнародного астрономічного союзу. Результати спостережень (координати на небосхилі, яскравість, температура, спектральний клас тощо) зберігаються у спеціальних базах даних. Відмінності параметрів одних і тих самих об'єктів можуть визначатися їх розташуванням, умовами та часом спостережень і мають усі підстави розглядатися як один з варіантів Big Data.

У зірковій астрономії виявлення зоряних кластерів та асоціацій, які можуть бути важливими для вивчення розвитку галактик і еволюції зірок, виконується засобами кластерного аналізу, як одним із методів статистичного аналізу, що дозволяє розподілити об'єкти на групи з важливими особливостями за певними критеріями.

II. Виклад основного матеріалу

Задача кластеризації даних є однією з важливих задач Data Mining, для розв'язання якої існує достатньо велика сукупність методів, зокрема викладених у [1-4]. Для виконання будь-якого з них необхідно мати дані про розміщення та раніше вказані характеристики зірок, кожна з яких може мати свою кількість градацій. Саме їх комбінації визначають певні важливі особливості об'єктів зоряного неба.

Методика кластерного аналізу полягає в розподілі зірок за вказаними параметрами, використовуючи певні критерії і послідовності дій. За допомогою відповідних статистичних алгоритмів зорі групуються у кластери з подібними характеристиками. Усередині кожної групи повинні виявитися «схожі» об'єкти, а об'єкти різних групи повинні бути якомога більш відмінні.

Мета дослідження – проаналізувати найбільш відомі класичні та нейромережеві методи кластеризації на базі характеристик об'єктів зоряного неба. Для дослідження обрані метод k-середніх, метод c-means, мережа Кохонена,

а також графічні методи (дендрограми), що дозволяють візуалізувати залежності між кластерами та зірковими об'єктами.

Метод k-середніх (k-means) розподіляє зірки на попередньо визначену кількість кластерів k. Для кожного з них у просторі ознак зірок випадковим чином обираються координати центру. Далі кожна точка даних з координатами об'єктів призначається до кластеру з найменшою відстанню до його центру, координати цього центру одразу перераховуються як середнє арифметичне координат всіх точок, які були попередньо призначені до цього кластеру. Такий процес повторюється до тих пір, поки кластери не стабілізуються. При цьому кількість ітерацій залежить від вірно вибраної кількості можливих кластерів, вдалого початкового вибору координат їх центрів і метрики для обчислення відстаней.

Метод c-means розділяє дані на кластери на основі подібності між ними, але кожна точка може належати до кількох кластерів з різною вагою. У цьому методі, що належить до групи методів нечіткої кластеризації, спочатку випадковим чином вибирається кількість кластерів, які називаються c. Кожна точка з координатами у просторі ознак об'єктів призначається до кожного кластеру з вагою, яка визначається відстанню між точкою та центроїдом кожного кластеру. Далі центроїди перераховуються на основі середньої ваги кожної точки, що належить до кластеру. Цей процес також повторюється до досягнення стабільності кластерів.

У нейромережевому підході розглянуто застосування мережі Кохонена [5]. Ця мережа із самоорганізацією може навчається без вчителя. При організації навчання для вимірювання відстані між вхідним вектором та вектором ваг кожного нейрона використана Евклідова метрика. При налаштуванні мережі було проварійовано кількість нейронів, розмір околиці та інші гіперпараметри.

A. Особливості застосування кластерного аналізу

Кластерний аналіз має одну суттєву особливість – він не є звичайним статистичним методом, оскільки до нього у більшості випадків незастосовні процеси перевірки статистичної значимості. Результати кластерного аналізу дають найбільш можливо-значиме рішення. Саме тому досить часто його використовують тоді, коли дослідник має набір даних, але не має виправданої апріорної гіпотези про класи цих даних.

Тому виділимо декілька застережень, які слід враховувати при використанні кластерниого аналізу:

- більшість методів кластерного аналізу є доволі таки простими евристичними процедурами, які, як правило, не мають класичного статистичного обґрунтування;
- різні методи кластеризації можуть породжувати різні кластерні рішення для одних і тих же даних. Це звичне явище у більшості прикладних досліджень, і тому слід по-перше обирати найбільш осмислене рішення, по-друге – завжди вказувати, який саме метод кластеризації було використано;
- дія кластерного аналізу полягає у привнесені структури у аналізовані дані. Тобто, кластеризація може призвести до появи артефактів (виявлення структури в даних, які її не мають);
- осмислене рішення при кластерному аналізі можна обрати лише тоді, коли є базис для цього – теоретичне обгрунтування. Без теоретичної моделі, без гіпотези стосовно структури даних з'являється небезпека наївного емпіризму, а саме прийняття результатів кластеризації за істину у кінцевій інстанції.

Усі розглянуті методи передбачають найпростішу форму розташування точок у просторі ознак кластеризації об'єктів зоряного неба, що визначено використанням Евклідової метрики.

B. Засоби моделювання та програмування

Для проведення дослідів і розрахунків, що виконані під час дослідження, використана мова програмування Python [6, 7]. В процесі роботи з мережею Кохонена використаний фреймворк Keras [8], який добре працює з Python, характеризується дружністю до користувачів і мінімалізмом, модульністю та легкою розширюваністю.

III. Висновки

Згадані методи опрацьовані на зображеннях зоряного неба, завдяки чому отримані порівняльні характеристики і проаналізовані обмеження.

Доведена ефективність кластеризації об'єктів зоряного неба за допомогою як методів k- та c-means, так і мережі Кохонена для обмеженого набору вхідних даних.

У подальшому описані методи планується застосувати до задач кластеризації зображень більшого обсягу спостережень об'єктів зоряного неба з доступних даних обсерваторій та порівняти їх результати між собою.

Література References

[1] Charu C. Aggarwal. Data Mining. The Textbook. Springer International Publishing. Switzerland, 2015. - 746p.

[2] B.S. Everitt, S.Landau, M. Leese, D.Stahl. Cluster Analysis. – London: Arnold Publishers, 2011. – 343p.

[3] A. K. Jain, R. C. Dubes. Algorithms for Clustering Data. – New Jersey:

[4] Thomas H. Cormen, Charles E. Leiserson, Ronald L. Rivest, Clifford Stein. Introduction to Algorithms, 4rd Edition. – The MIT Press, 2022. – 1312p.

[5] Нейронна мережа Кохонена. Вінницький національний технічний університет. https://inmad.vntu.edu.ua/portal/static/D28C6207-9C44-4C8C-B56E-43E2CF72C372.pdf

[6] Wes McKinney. Python for Data Analysis, 3rd Edition - USA. :O'Reilly Media, Inc. – 2022. – 342 p.

[7] Jake VanderPlas. Python Data Science Handbook, 2nd Edition - USA. : O'Reilly Media, Inc. - 2022. – 342 p.

[8] Keras. Інститут теоретичної фізики ім. М.М. Боголюбова. http://cloud-5.bitp.kiev.ua/?page_id=600