УДК 681.518.54

Харчук В.В., аспірант спеціальності 123 Комп'ютерна Інженерія
Науковий керівник: Ткаченко С.М., доцент кафедри інформаційних технологій та комп'ютерної інженерії
(Національний технічний університет «Дніпровська політехніка», м. Дніпро, Україна)

КЛАСТЕРНИЙ АНАЛІЗ ДАНИХ МАШИННОГО ЗОРУ ДЛЯ МОДЕЛЮВАННЯ ПРОСТОРУ

Кластеризація — це техніка машинного навчання та аналізу даних, яка передбачає групування подібних точок даних на основі певних особливостей або характеристик. Це тип алгоритму неконтрольованого навчання в машинному навчанні.

Мета кластеризації полягає в тому, щоб розділити набір даних на групи або кластери таким чином, щоб точки даних у кластері були більш схожі одна на одну, ніж на точки в інших кластерах. Це допомагає визначити закономірності, структури або зв'язки в даних.

Кластерний аналіз даних машинного зору - це процес групування великої кількості візуальних даних у "кластери" або групи, що мають спільні ознаки. Це допомагає в розумінні структури даних та виявленні закономірностей або складних взаємозв'язків [2].

Основні ознаки:

Неконтрольоване навчання: кластеризація зазвичай є завданням неконтрольованого навчання, тобто алгоритм не покладається на позначені дані з попередньо визначеними категоріями. Натомість він визначає шаблони або групування виключно на основі внутрішньої структури даних.

Метрика подібності: алгоритми кластеризації використовують метрику подібності, щоб визначити, наскільки близькі або схожі точки даних одна до одної. Вибір показника подібності залежить від характеру даних і конкретних вимог до аналізу[1].

Мета: основна мета кластеризації полягає в тому, щоб максимізувати внутрішньокластерну подібність і мінімізувати міжкластерну подібність. Іншими словами, точки в одному кластері мають бути більш схожими, а точки в різних кластерах мають бути менш схожими.

Це популярний алгоритм кластеризації на основі центроїда, який розбиває дані на певну кількість (k) кластерів. Він спрямований на мінімізацію суми квадратів відстаней між точками даних і центроїдом призначеного їм кластера. K-means залишається одним із найпоширеніших алгоритмів кластеризації завдяки своїй простоті та ефективності.

Щоб виконати аналіз кластеризації за допомогою K-середніх, нам потрібно використати техніку під назвою «Аналіз основних компонентів» (PCA), щоб вибрати найкращі функції/компоненти для використання.

Аналіз головних компонентів, або PCA, — це метод зменшення розмірності, який часто використовують для зменшення розмірності великих наборів даних шляхом перетворення великого набору змінних у менший, який все ще містить більшу частину інформації у великому наборі.

В ідеалі в нашому наборі даних ми зазвичай маємо більше 2 функцій, PCA дає нам змогу вибрати найбільш релевантні функції для кластерного аналізу. Наприклад, якщо ви хочете визначити різні кластери клієнтів у магазині, у вашому наборі даних ви матимете багато характеристик конкретного клієнта. Тип автомобіля, на якому вони їздять, може бути доречним, щоб показати, наскільки вони багаті, але колір автомобіля не має значення, тип продуктів, які вони купують, важливий, але їхній зріст може не мати

значення. Завдяки PCA ми зможемо визначити важливі характеристики, які можуть допомогти нашим кластерам.

Функція вилучення використовується для отримання необроблених даних за допомогою методів камерного бачення, таких як SPP-net і GMM, щоб керувати інтелектуальними даними про дорожній рух у режимі реального часу та ідентифікувати дані про затори на шосе. Іноді зібрані дані можуть бути простими або складними. Відповідно, як виділення функцій, так і дані датчиків переднього плану передаються на наступний рівень на основі представлення даних датчиків, тобто аналізу даних датчиків Інтернету речей для продуктивності та інтерпретації відповідних даних за допомогою Simulink і методів інтеграції даних[3].

Таким чином, аналіз даних доступний у всьому IoT. У наш час генерується величезна кількість різноманітних даних. Таким чином, отримання даних датчика та виконання аналізу отриманих даних є складним завданням. У цій роботі автори представили підхід MSDACA для збору та аналізу датчиків IoT. Хмарна платформа ThingSpeak IoT із кодом Matlab реалізувала запропоновану структуру. Також розроблено модель Simulink для аналізу даних і візуалізації отриманої інформації в режимі реального часу.

Список використаних джерел:

1. Hans-Dieter W. Machine Learning, Deep Learning, and AI: What’s the Difference? – 2017. – C. 2–4. URL: https://www.researchgate.net/publication/318900216_Machine_Learning_Deep_Learning_and_AI_What's_the_Difference

2. Pattern recognition guide. URL: https://www.v7labs.com/blog/pattern-recognition-guide

3. Chaves M. GLCMs — a Great Tool for Your ML Arsenal. – 2022. URL: https://towardsdatascience.com/glcms-a-great-tool-for-your-ml-arsenal-7a59f1e45b65