Кластеризація. Модель k-means

Актуальність	Дев'яносто відсотків даних у світі сьогодні створено лише за останні два роки. Щодня створюється близько трьох мільярдів гігабайтів нової інформації. Збереження такого об'єму даних потребує немалих ресурсів та зусиль.Але найбільша проблема полягає не в збереженні даних, а в їх обробці – за прогнозами в 2020 році лише 35% даних будуть корисними. Тому інтелектуальний аналіз даний є актуальною проблемою сьогодення. Кластерний аналіз є потужним механізм для спрощення даних для подальшого їх аналізу
Постановка проблеми	Сьогодні використовуються різні модифікації даного алгоритму, але кожен із них має два основних недоліки: необхідність знати кількість кластерів перед виконанням алгоритму і оптимальність отриманих рішення не є гарантованою.
Шляхи вирішення проблеми	Тому наша модифікація буде направлена на прискорення роботи алгоритму за допомогою використання паралельних обчислень. Розроблений алгоритм представляє собою ітераційну процедуру: - Крок 1 (CPU). Випадковим чином обрати центри кластерів із елементів вхідних даних, встановити номер ітерації $l = 0$. - Крок 2 (CPU). Підключити GPU модуль, обрахувати конфігурацію запуску для функції-ядра:
Результати та висновок	В даному розділі проведено аналіз алгоритму k-means на предмет можливості розпаралелювання обчислень. Виявлено, що в цьому алгоритмі операції обчислення матриці приналежності та центрів кластерів можна виконати паралельно. Наведено модифіковані алгоритми кластеризації для масово паралельних обчислень на графічних процесорів Nvidia. Математичний аналіз складності алгоритму показує велике теоретичне прискорення для розроблених модифікацій алгоритмів в порівнянні з виконанням на центральному процесорі, що може досягати більше ста для модифікованого алгоритму k-means.