Звада І. Р., студент групи ІПЗм-20-2,
Марчук Г. В., ст. викл. каф. КН
Державний університет «Житомирська політехніка»

ОБГРУНТУВАННЯ ЗАСТОСУВАННЯ АЛГОРИТМУ КЛАСТЕРИЗАЦІЇ DBSCAN В СИСТЕМІ ДОСТАВКИ

Кластерний аналіз або просто кластеризація – це, по суті, метод нав-чання без вчителя, який ділить точки даних на ряд певних груп, так що точки даних в одних і тих же групах мають схожі властивості, а точки даних в різних групах мають різні властивості. Кластерний аналіз складається з безлічі різних методів, заснованих на різній еволюції.

Наприклад, K-середні, метод Варда, DBSCAN, спектральна кластеризація тощо. По суті, всі методи кластеризації використовують один і той же підхід, тобто спочатку знаходяться подібності, а потім вони використовуються для кластеризації даних в групи.

У 2014 році алгоритм DBSCAN отримав премію «перевірено часом» (премія дається алгоритмам, які отримали значну увагу в теорії і практиці) на провідній конференції з інтелектуального аналізу даних [1].

DBSCAN (Density-based spatial clustering of applications with noise) запропонували М. Естер, Г. Кригель, Y.Сандер і С. Су в 1996 році. Алгоритм слід використовувати для пошуку асоціацій і структур в даних, які важко знайти вручну, але які можуть бути актуальними і корисними для пошуку закономірностей і прогнозування тенденцій.

На рисунку 1 показано набір точок. Навколо кожної точки проведено круговий кордон з радіусом R у вигляді пунктирною лінії. Щоб проілюструвати концепцію щільності, алгоритм визначає 3 типи точок:

– Основні точки - точки, для яких існує не менше N точок, включаючи її саму, в межах власного кордону. Нехай N = 3, тоді всі точки з червоним кордоном – основні. Основні точки досяжні з усіх інших основних точок в межах їх кордонів. З'єднуємо їх двобічної стрілкою.

– Неосновні, але досяжні - це точки, в межах яких менше N точок, включаючи її саму. На рисунку 2 при N = 3 всі точки з синім кордоном є не основними. Якщо неосновна точка має центральну точку в межах своєї межі, вона вважається досяжною з цієї базової точки. За визначенням, основна точка не може бути досягнута з неосновної точки. Синя точка (рис.2) є не основною, але досяжною.

– Викиди. Неосновні точки, недоступні ні для яких точок. Помаранчева точка (рис.2) – викид.

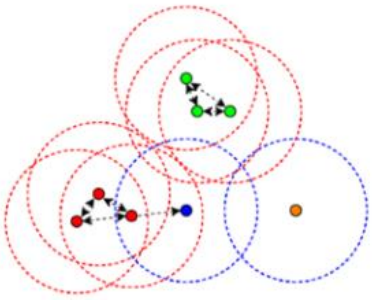

Рис. 1. Ілюстрація алгоритму DBSCAN

Алгоритмічні кроки для кластеризації DBSCAN (рис.2):

– Алгоритм діє шляхом довільного вибору точки в наборі даних, Інформація про її сусідів витягується з параметра ε;

– Якщо є точки minPts в радіусі ε, починається формування кластера;

– Якщо точка виявляється центральною, то точки в околиці points також є частиною кластера;

– Процес триває до тих пір, поки щільно пов'язаний кластер не буде повністю знайдений.

Перевагами даного методу є те, що не потрібно вказувати кількість кластерів, також метод може знаходити кластери будь-яких форм. Також одною з найбільш важливих переваг є наявність поняття шуму. Ще одною перевагою є те, що метод потребує лише два параметри і в переважній більшості випадків є нечутливим до впорядкування точок.

Недоліком є те, що в ситуації, коли точка находиться в межі досяжності більше ніж одного кластеру, то вона може належати одному з кластерів в залежності від порядку обробки даних. Також недоліком є те, що якість кластеризації залежить від функції відстані. Також нерентабельно застосовувати алгоритм для даних з великою щільністю.

Отже найбільш вдалим вибором для системи доставки буде метод DBSCAN, так як одною із основних переваг є стійкість до викидів, а їх наявність в подібних системах є доволі розповсюдженим явищем.

Список літератури

1. 2014 SIGKDD TEST OF TIME AWARD, 2014. https://www.kdd.org/news/view/2014-sigkdd-test-of-time-award
2.M. Ester, H-P. Kriegel, J. Sander, X. Xu. A density-based algorithm for discovering clusters in large spatial databases with noise // Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96) / Evangelos Simoudis, Jiawei Han, Usama M. Fayyad. - AAAI Press.