МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
ЧОРНОМОРСЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ ІМЕНІ ПЕТРА МОГИЛИ

Паленко Роман Олегович

УДК 004.01

СИСТЕМА КЛАСТЕРИЗАЦІЇ АВТОМОБІЛЬНИХ ДОРІГ З ВИКОРИСТАННЯМ ІНТЕЛЕКТУАЛЬНОГО АНАЛІЗУ ДАНИХ

Галузь знань 12 «Інформаційні технології» за спеціальністю
122 «Комп’ютерні науки та інформаційні технології»
122 - ДР.А - 403.21610316

Автореферат
дипломної роботи на здобуття освітньої кваліфікації
«бакалавр комп’ютерних наук та інформаційних технологій»

Миколаїв – 2020

Дипломною роботою є рукопис.

Робота виконана в Чорноморському національному університеті імені Петра Могили Міністерства освіти і науки України на кафедрі інтелектуальних інформаційних систем

Науковий керівник: к.т.н, доцент, доцент кафедри ІІС
Сіденко Євген Вікторович

Рецензент: к.ф.-м.н., доцент, доцент кафедри ІІС,
секція прикладної та вищої математики
Воробйова Алла Іванівна

Захист відбудеться «24» червня 2020 р. о 9^{30} год. на засіданні екзаменаційної комісії (ауд. 2-406) у Чорноморському національному університеті імені Петра Могили за адресою: 54003, м. Миколаїв, вул. 68 Десантників, 10.

З дипломною роботою можна ознайомитися в бібліотеці Чорноморського національного університету імені Петра Могили за адресою: 54003, м. Миколаїв, вул. 68 Десантників, 10.

Автореферат представлений «15» червня 2020 р.

Секретар
екзаменаційної комісії,
викладач кафедри ІІС О. С. Скакодуб

ЗАГАЛЬНА ХАРАКТЕРИСТИКА РОБОТИ

На сьогоднішній день кожен мешканець України щодня користується дорогами пересуваючись на роботу, у відпустку чи просто до магазину. І більшість автошляхів якими користуються пересічний мешканець країни у досить поганому стані.

Зараз можна сказати, що в нашій країні геть погане дорожнє покриття, але це не так. Кожен день виконуються ремонтні роботи на кілометрах автошляхів. Але через хаотичність таких робіт та розпорошеність по території нашої величезної країни вони стають непоміченими.

Основною проблемою є вибір пріоритетів проведення ремонтних робіт. Одна дорога може належати декільком областям, які матимуть різні локальні пріоритети щодо її ремонту. Це ставить нас користувачів у скрутне положення, коли частина одного автошляху у відмінному стані, а інша може бути у після воєнному стані.

Актуальність теми дипломної роботи полягає у тому, що вона дозволить проводити кластеризацію великого обсягу даних автошляхів України. Правильний кластерний аналіх дозволить прийняти правильне рішення щодо проведення у майбутньому ремонтних робіт.

Метою дипломної роботи є програмний додаток, що виконуватиме кластеризацію автодоріг за визначеними атрибутами. Користувач не матиме проблем з величезною купою розрахунків, йому лише необхідно створити вибірку з даними, а додаток самотужки проведе процес кластеризації. Інтерфейс повинен бути зручним та зрозумілим, щоб будь-яка особа без проблем могла використовувати даний застосунок. Додаток повинен бути адаптивним, тобто працювати з різними наборами даних, що забезпечить його актуальність.

Об'єктом даної роботи є програмний додаток з кластеризації автошляхів

Предметом принципи та засоби кластеризації даних за обраним алгоритмом.

Практичне значення застосунку полягає у можливості застосування отриманого застосунку для кластеризації даних методом C-means, а також надання зрозумілого та зручного інтерфейсу для взаємодії користувачів з застосунком

Дипломна робота складається зі вступу, 4 розділів, висновків, переліку джерел посилання та додатків. Загальний обсяг роботи складає __ сторінок (без додатків), __ рис., __ табл., __ додатки та __ джерел посилання на літературні джерела.

ОСНОВНИЙ ЗМІСТ РОБОТИ

У вступі дипломної роботи обґрунтовано актуальність обраної теми, проблеми які вирішує дипломна робота, сформульовано мету і задачі дослідження, визначено предмет та об’єкт дослідження.

У першому розділі проведено дослідження з існуючих систем кластеризації автошляхів Украйни. Виокремлено основні види автошляхів, проблематику класифікацій та їх суперечність. З досліджених видів класифікації було зроблено висновок про їх повноцінне покриття всіх характеристик автошляхів України. Проте було виявлено, що користуватися ними досить незручно для прийняття рішень, щодо ремонту або кластеризації їх, через те, що ці класифікації майже не корелюють між собою. Було проведено аналіз класичних алгоритмів кластеризацій даних, а також метрик відстаней які вони використовують. Особливу увагу було надано алгоритмат K-means та C-means. Серед них було обрано другий, як найбільш підходящий. Він має декілька переваг над першим, та дозволяє проводити не жорству кластеризацію надаючи більше даних користувачу для подальшого прийняття рішення.

Найбільш популярним алгоритмом нечіткої кластеризації є алгоритм c-середніх. Він є модифікацією метода k-середніх, там має наступні основні кроки алгоритму:

1. Обрати початкове нечітке розбиття n об’єктів на k кластерів шляхом створення матриці належності U розміру n на k;
2. Використовуючи матрицю U, знайти значення критерія нечіткої помилки;
3. Перегрупувати об’єкти з метою зменшення значення критерію нечіткої помилки;
4. Повертатися до другого кроку доки зміни матриці U не стануть незначними.

Недоліки та переваги алгоритму c-середніх аналогічні алгоритму k-середніх, але за рахунок нечіткого відношення об’єктів до кластерів, цей метод дозволяє

визначати приналежність об’єктів, що знаходяться на кордонах кластерів. Нечіткий алгоритм с-середніх був представлений Данном у 1973 році, а далі покращений Бездеком у 1981 році

У другому розділі було виконано дослідження структури вибірки даних для вирішення поставленої задачі. Було розглянуто проблематику створення вибірок та шляхи їх подолання. Були використані доступні офіційні джерела даних для побудови базової вибірки. Нажаль цих даних виявилося недостатньо. Детальний огляд інформації з відкритих джерел виявив, що досить велика кількість інформації не опублікована. Так навіть в офіційних звітах Автодору про витрати коштів на ремонт доріг не зазначається на які автошляхи виділяються кошти, а лише надається інформація в якому обсязі ці кошти отримують області. А також було проведено аналіз побічних характеристик автошляхів для виконання повного кластерного аналізу, дані були отримані аналітичним шляхом.

Для відображення характеристики класифікації за значенням достатньо перевести кожну класифікацію в числову характеристику з кроком в умовну одиницю, таблиця 1

Таблиця 1

Переведення класифікацій доріг в умовні позначення

Назва класу дороги	Кількість умовних одиниць
Міжнародні	6
Національні	5
Регіональні	4
Територіальні	3
Районні	2
Сільські	1

На щастя, велику кількість інформації можна віднайти за допомогою аналізу, зокрема важливу характеристику про важливість автошляху для населення було б доцільно виразити у кількості прибутку, що несе цей автошлях для країни сполучаючи транспортні вузли, але не маючи даної інформації у відкритому доступі довелося перетворити цю характеристику у кількість великих населених пунктів.

При чому довелося робити додаткове дослідження з визначення поняття «великий населений пункт».

Перш за все нас повинна цікавити інформація про населені пункти, що поєднує той чи інший автошлях. Першим і найпростішим рішенням було б підрахувати кількість населених пунктів сполучених дорогою та зробити це значення характеристикою для нашої вибірки. Але відразу помітні підводне каміння у вигляді нерівнозначності різних міст, як ми можемо рахувати Київ – найбільше місто України за один населений пункт і в той же час у відповідність йому ставити селище в якому проживають 28 осіб. Вийде ситуація в якій одна дорога матиме 7-8 населених пунктів, але це будуть великі обласні та районні центри, а інша 20-30 населених пунктів, во вздовж неї розклалися десятки сіл та містечок. Тому потрібно вводити альтернативний варіант числення. Було розглянута два варіанти:

Перший підхід визначав умовну одиницю числення за кожну тисячу або п’ять жителів населеного пунктів. Для кожної дороги ці показники додавались і були отримані певні значення.

Другий підхід залишав значенням атрибуту кількість населених пунктів, але на відміну від початкового варіанту, він передбачав певну валідацію даних, тобто населені пункти з кількістю жителів меншою за певний поріг не брались до розрахунку.

З цих двох підходів було обрано другий. Перший повністю вирішував проблему «нечесної» оцінки населених пунктів, але створював проблему неструктурованості даних – відносно коротка дорога, що поєднує Київ та Харків, за цією системою буде мати більше умовних одиниць ніж довга дорога через всю країну що оминає обласні центри. Межею для другого підходу було обрано 50 тисяч населення. Ця межа була обрана шляхом спостережень за чисельністю міст на головних магістральних шляхах країни. Загалом якщо розглядати всі населені пункти, що були досліджені під час роботи над проектом можна виділити чотири групи: обласні міста з населенням більше декількох сотень тисяч, міста з населенням від 50 до 100 тисяч населення, містечка та селища з населенням до 15 тисяч, та невеличкі села з населенням до 1 тисячі мешканців. Тому межа у 50 тисяч

виглядає розумним обмеженням для виокремлення саме великих та важливих населених пунктів. Щоправда до цієї системи довелося додати одне виключення – деякі автошляхи закінчуються на кордоні України у контрольно пропускних пунктах та поселеннях. Вони звичайно і близько не мають населення у 50 тисяч мешканців, але є важливими населеними пунктами для населення що подорожує в інші країни, та для України в цілому, бо через них виконується імпорт та експорт продукції. Тому такі місця на карті також рахувалися при виконанні роботи за «великий населений пункт».

Було детально розглянуто обраній алгоритм кластеризації даних – C-means. Та обрано для нього підходящу метрику.

В третьому розділі був проведений розгляд реалізації алгоритму. Зроблені акценти на основні частини коду, пояснення обраних алгоритмів та засобів для створення застосунку. Була проведена робота з представлення свого рішення окремих проблем. При реалізації алгоритму було враховано слабкі сторони алгоритму, та можливість роботи з різними даними. Було проведена робота зі створення системи перевірки даних, які надає користувач, для правильної роботи застосунку. Створені системи сповіщення користувача про проблеми, що виникли в ході роботи з програмою. Користувацький інтерфейс застосунку простий та лаконічний. Він дозволяє просто та зручно взаємодіяти з резільтатами кластеризації.

розглянемо функціонал інтерфейсу, а далі вже детально пройдемо по основним моментам. Користувач може взаємодіяти з шести кнопками та двома комбобоксами. Рис. 1.

Зміст кнопок:

1. Load – відповідає за вибір користувачем файлу з даними та зчитуванням з файлу;
2. Clusterize – відповідає за проведення процесу кластеризації, збереження результатів в буфері програми, наповнення комбобоксів;
3. Save – відповідає за запис результатів кластеризації у файл;

4. Кнопки «<» та «>» – відповідають за вибір ітерації кластеризації для побудови графіку;
5. Build – відповідає за побудову графіку за вибраними атрибутами у комбобоксах;
6. Комбобокси – відповідають за вибір атрибутів для побудови графіку.

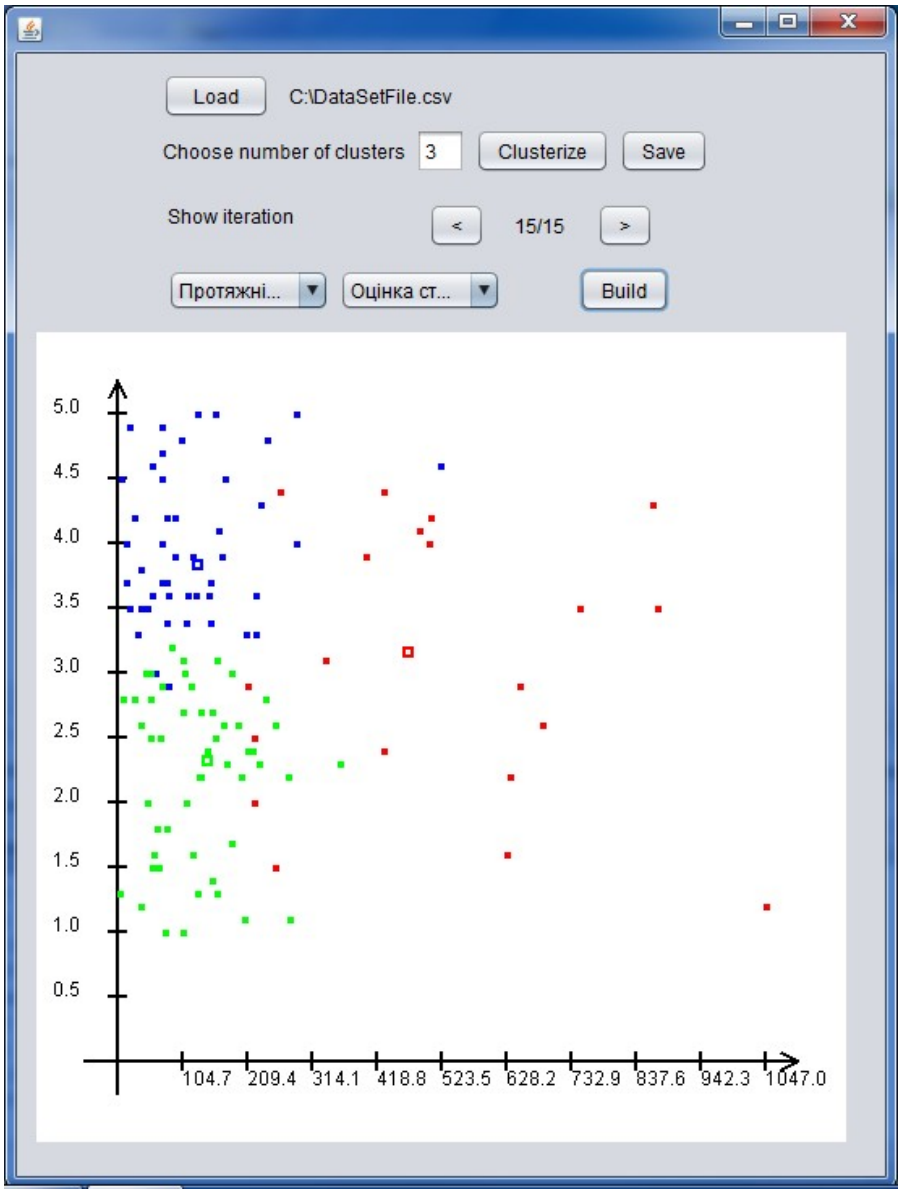

Рис. 1. Скріншот інтерфейсу програми

В четвертому розділі був проведений інструктаж для користувача з використання програмного застосунку. Було розглянуто структуру вибірки, що потрібно надавати програмі для коректної роботи, та зауваження до основних

елементів інтерфейсу. Розяснено як процювати з помилками, що виникають при роботі та можливих шляхів їх подолання.

У розділі з охорони праці роботи були викладені вимоги до робочого місця інженера-програміста. Створені умови повинні забезпечувати комфортну роботу. На підставі вивчення літератури з цієї теми було визначено оптимальні розміри робочого столу і крісла, робочої поверхні, а також проведено вибір системи і розрахунок вентиляційної системи виробничого приміщення. Дотримання умов визначає оптимальну організацію робочого місця інженера-програміста, що дозволить зберегти максимальну працездатність протягом всього робочого дня, підвищить, як у кількісному, так і в якісному відношенні, продуктивність праці програміста, що у свою чергу сприятиме швидкій розробці та налагодженню програмного продукту.

ЗАГАЛЬНІ ВИСНОВКИ

В ході роботи на дипломним проектом було детально досліджено актуальну систему класифікації автошляхів України. Вона являє собою дещо заплутану систему класифікацій що не жорстко корелюють між собою, а в певному плані навіть дублюють ті чи інші характеристики автодоріг.

Було виявлено проблему інформування населення про теперішній стан автомобільних шляхів владою та відсутність навіть базової інформації у відкритих джерелах. Через це була проведена аналітична робота за збору інформації про різні характеристики доріг для подальшого процесу кластеризації.

Було розглянуто різні класичні методи кластеризації даних. Серед них шляхом оцінки їх доцільності для поставленої задачі був обраний один, а саме алгоритм нечіткої кластеризації c-середніх.

Було спроектовано структуру обраного алгоритму, розглянуто різні метрики відстані, що могли використовуватися в процесі кластеризації.

Був створений програмний застосунок з реалізованим алгоритмом кластеризації, зручним користувацьким інтерфейсом та візуалізацією даних.

Теоретична значимість роботи полягає у можливості використання отриманих результатів для проведення кластеризації даних, та оцінки якості інформаційного забезпечення населення з поставленої задачі.

Практична значимість полягає у можливості застосування отриманого застосунку для кластеризації даних методом C-means, а також надання зрозумілого та зручного інтерфейсу для взаємодії користувачів з застосунком.

Дослідження й виявлення можливих причин виробничих нещасних випадків, професійних захворювань, аварій, вибухів, пожеж, і розробка заходів і вимог, спрямованих на усунення цих причин дозволяють створити безпечні й сприятливі умови для праці людини. Комфортні й безпечні умови праці − один з основних факторів, який впливає на продуктивність і безпеку праці, здоров’я працівників.

Під час роботи над дипломною роботою не було виявлено жодних порушень з питань охорони праці. Робоче місце було оснащене належним чином. Технічний

стан обладнання відповідав стандартам безпеки і нормам охорони праці, ніяких дефектів обладнання під час виконання роботи не виявлено.

В результаті написання спеціальної частини з охорони праці було досягнуто поставленої мети, а саме створення безпечних і здорових умов праці на робочих місцях, в робочих зонах, у виробничих приміщеннях

АНОТАЦІЯ

бакалаврської дипломної роботи Паленка Романа Олеговича
на тему: «Система кластеризації автомобільних доріг з використанням інтелектуального аналізу даних»

Дипломна робота присвячена питанню дослідження і аналізу існуючих методів класифікації автомобільних доріг та проблеми їх кластеризації, для подальшого впровадження в програмний застосунок.

В роботі описано основні принципи та методи проведення кластерного аналізу даних. Спираючись на існуючі системи класифікації автомобільних доріг була розроблена вибірка даних для подальшого кластерного аналізу. Створений програмний застосунок повністю виконує поставлену задачу та надає користувачу результат не тільки в текстовому, а й в графічному форматі.

Фахова частина включає вступ, чотири розділи, висновки та додатки до дипломної роботи. Спеціальна частина включає розділ про охорону праці та безпеку на підприємстві.

В першому розділі розглядаються існуюча система класифікації автодоріг, проблеми що виникають при використанні цих класифікації, розглянуті класичні алгоритми кластеризації даних, та обрано найбільш підходящий для виконання поставленої задачі.

В другому розділі проводиться аналіз проблематики формування вибірки даних, розглядаються джерела інформації для формування вибірки та проводиться аналітичне створення вибірки.

У третьому розділі розглядається програмна реалізація обраного алгоритму, використаних методів для досягнення мети.

В четвертому розділі розглядається інструкція користувача з використання створеного застосунку.

Дипломна робота містить: сторінок –96, рисунків – 21, таблиць – 18, додатків –2, джерел – 34.

Ключові слова: *кластеризація, вибірка, автошлях, класифікація*.

ABSTRACT

graduate work on the topic: "Highway clustering system using data mining"

student of group 403 Roman Palenko

Graduate work is devoted to the study and analysis of existing methods of classification of roads and the problem of their clustering, for further implementation in the software application.

The work describes the basic principles and methods of cluster data analysis. Based on the existing road classification systems, a data sample was developed for further cluster analysis. The created software application completely fulfills the set task and gives the user the result not only in text, but also in graphic format.

The professional part includes an introduction, four sections, conclusions and appendices to the graduate work. The special part includes a section on labour protection and safety at the enterprise.

The first section discusses the existing system of road classification, the problems that arise when using these classifications, considers the classical algorithms of data clustering, and selected the most suitable for the task.

The second section analyzes the problems of data sampling, considers the sources of information for sampling and analytical creation of the sample.

The third section discusses the software implementation of the selected algorithm, the methods used to achieve the goal.

The fourth section discusses the task of informing the user about the use of the created application.

Thesis contains: pages - 96, figures - 21, tables - 18, appendices - 2, sources - 34.

Keywords: *clustering, sampling, highway, classification*.