НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ УКРАЇНИ
«КИЇВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ
імені ІГОРЯ СІКОРСЬКОГО»

ДИПЛОМНА РОБОТА НА ТЕМУ:

Методи кластеризації даних в пакеті Orange та їх узагальнення

Виконав:
студент гр. КА-44
Буханенко Андрій Олександрович
Науковий керівник:
професор, д.ф.м.н. Данилов Валерій Якович

Об'єкт дослідження: кластерний аналіз в пакеті Orange.

Предмет дослідження: узагальнення методів кластеризації.

Мета дипломної роботи: розробити порівняльну характеристику метдів кластеризації, створити методичне забезпечення для програмного забезпечення Orange.

Програмні засоби: програмне забезпечення Orange.

Результати: розроблено порівняльну характеристику метдів кластеризації, створено методичне забезпечення для програмного забезпечення Orange.

Напрямок подальших досліджень: збільшити кількість доступних методів кластеризації в пакеті Orange, вдосконалити на розширити методичне забезпечення для Orange.

Методи кластерного аналізу дозволяють вирішувати такі завдання:

• проведення класифікації об'єктів з урахуванням ознак, що відображають сутність, природу об'єктів. Рішення такого завдання, як правило, призводить до поглиблення знань про сукупність об'єктів, що класифікуються;

• перевірка висунутих припущень про наявність певної структури у досліджуваній сукупності об'єктів, тобто пошук існуючої структури;
побудова нових класифікацій для недостатньо вивчених явищ, коли необхідно встановити наявність зв'язків усередині сукупності і спробувати принести в неї структуру.

Обґрунтуванням актуальності

• вирішення задачі кластеризації є необхідність усунення усереднення при ідентифікації невідомих залежностей, оскільки в цьому випадку прогнозні оцінки є зміщеними, і їх не можна вважати адекватними. Тому ідентифікацію необхідно здійснювати в класі однотипних об'єктів або процесів, передумовою чого є попередня кластеризація.

Міра відстаней

1. Евклідова відстань

$$\rho(x,x') = \sqrt{\sum_{i}^{n}(x_i - x_i')^2}$$

2. Квадрат евклідової відстані

$$\rho(x,x') = \sum_{i}^{n}(x_i - x_i')^2$$

3. Відстань міських кварталів (Манхеттенська відстань)

$$\rho(x,x') = \sum_{i}^{n}|x_i - x_i'|$$

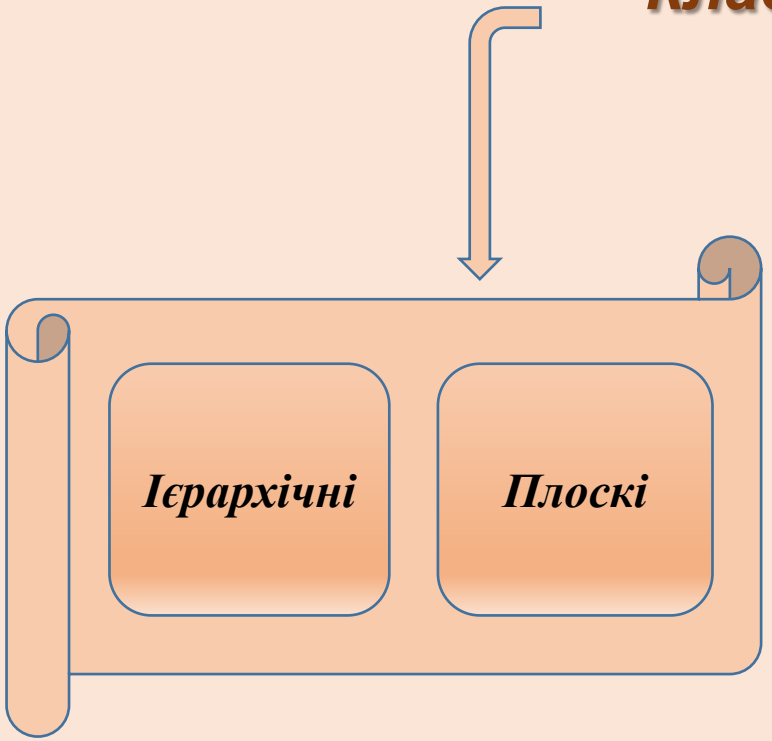
4. Відстань Чебишева

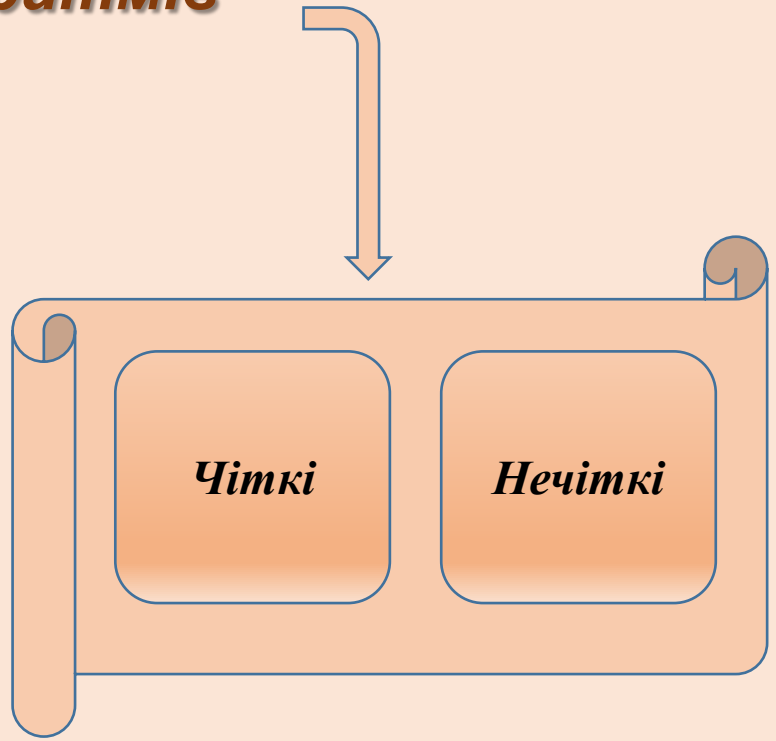
$$\rho(x,x') = \max(|x_i - x_i'|)$$

5. Відстань степіні

$$\rho(x,x') = \sqrt[r]{\sum_{i}^{n}(x_i - x_i')^p}$$

Класифікація алгоритмів





Розглянуті методи клатеризації

1. Алгоритми ієрархічної кластеризації

2. Алгоритми квадратичної помилки

3. Нечіткі алгоритми

4. Алгоритми, засновані на теорії графів

5. Алгоритм виділення зв'язкових компонент

6. Алгоритм мінімального покривного дерева

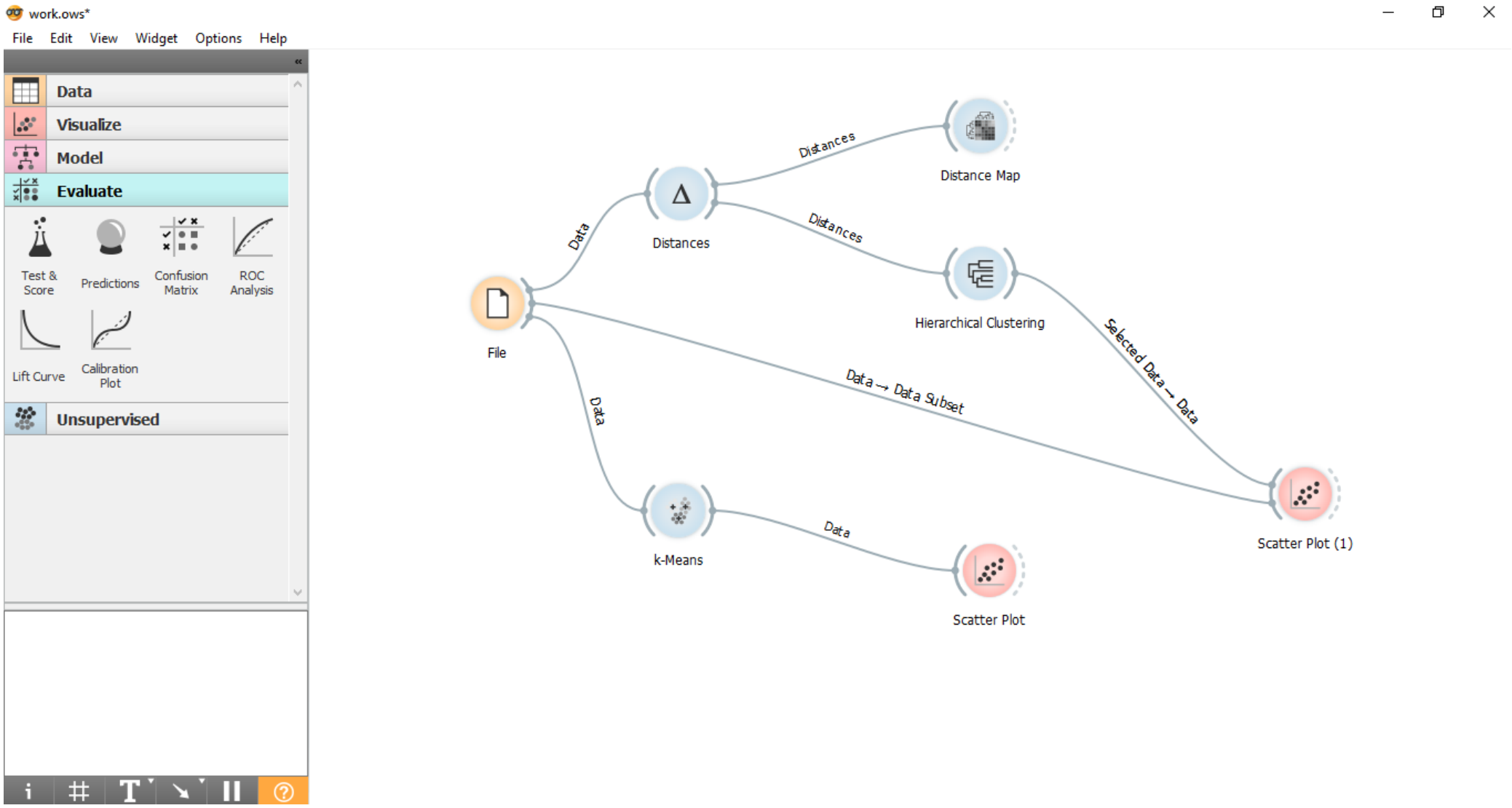
7. Пошарова кластеризація

Обчислювальна складність алгоритмів

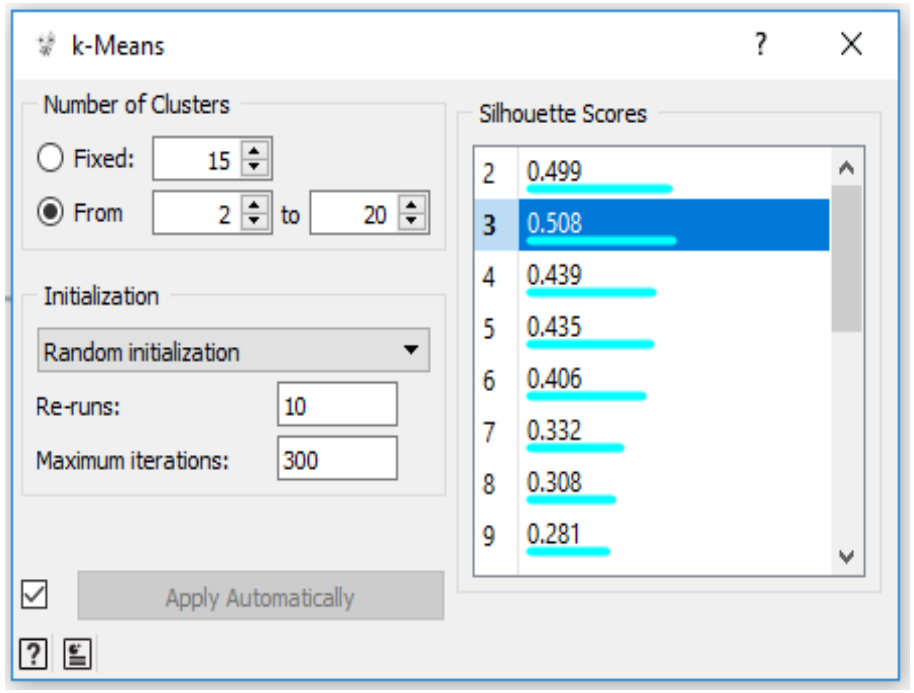
Алгоритм	Обчислювальна складність
Ієрархічний O (n2)	O (n^2)
k-середніх	O (nkl), де k - число кластерів, l - число ітерацій
c-середніх	
Виділення зв'язкових компонент	залежить від алгоритму
Мінімальна покривне дерево	O (n^2 log n)
Пошарова кластеризація	O (max (n, m)), де m <n (n-1) / 2

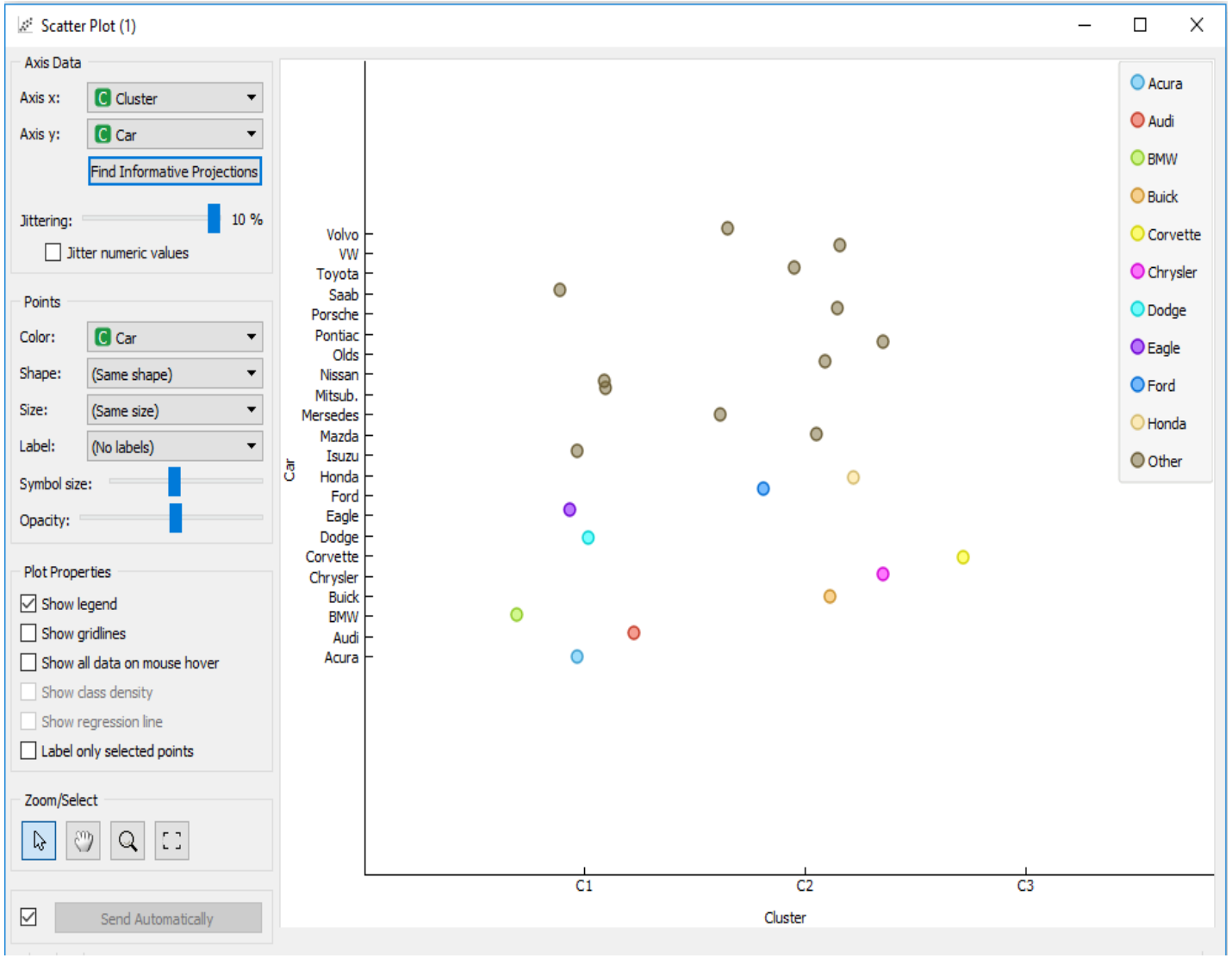
Порівняльна таблиця алгоритмів

Алгоритм	Форма кластерів	Вхідні дані	Результати
Ієрархічний	Довільна	Число кластерів або поріг відстані для усічення ієрархії	Бінарне дерево кластерів
k-середніх	Гіперсфера	Число кластерів	Центри кластерів
c-середніх	Гіперсфера	Число кластерів, ступінь нечіткості	Центри кластерів, матриця приналежності
Виділення зв'язкових	Зв'язкових компонент	Поріг відстані R	Деревовидна структура кластерів
Мінімальна покриває дерево	Довільна	Число кластерів або поріг відстані для видалення ребер	Деревовидна структура кластерів
Пошарова кластеризація	Довільна	Послідовність порогів відстані	відстані Деревовидна структура кластерів з різними рівнями ієрархії

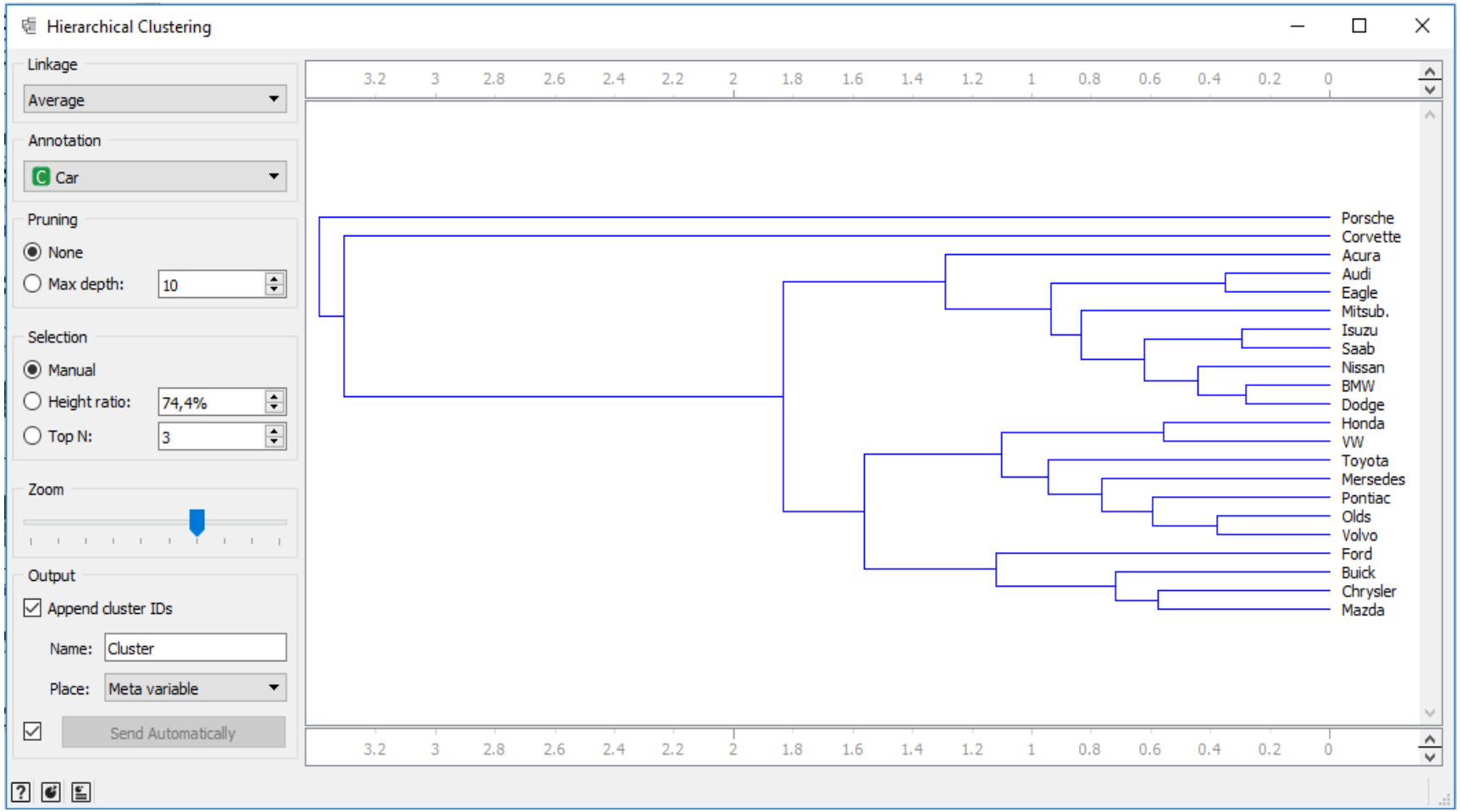

work.ows*
File Edit View Widget Options Help
Data
Visualize
Model
Evaluate
Test & Score
Predictions
Confusion Matrix
ROC Analysis
Lift Curve
Calibration Plot
Unsupervised
Distances
Distance Map
Distances
Data
Distances
Hierarchical Clustering
File
Selected Data → Data
Data → Data Subset
Data
k-Means
Data
Scatter Plot
Scatter Plot (1)

«k-Means»

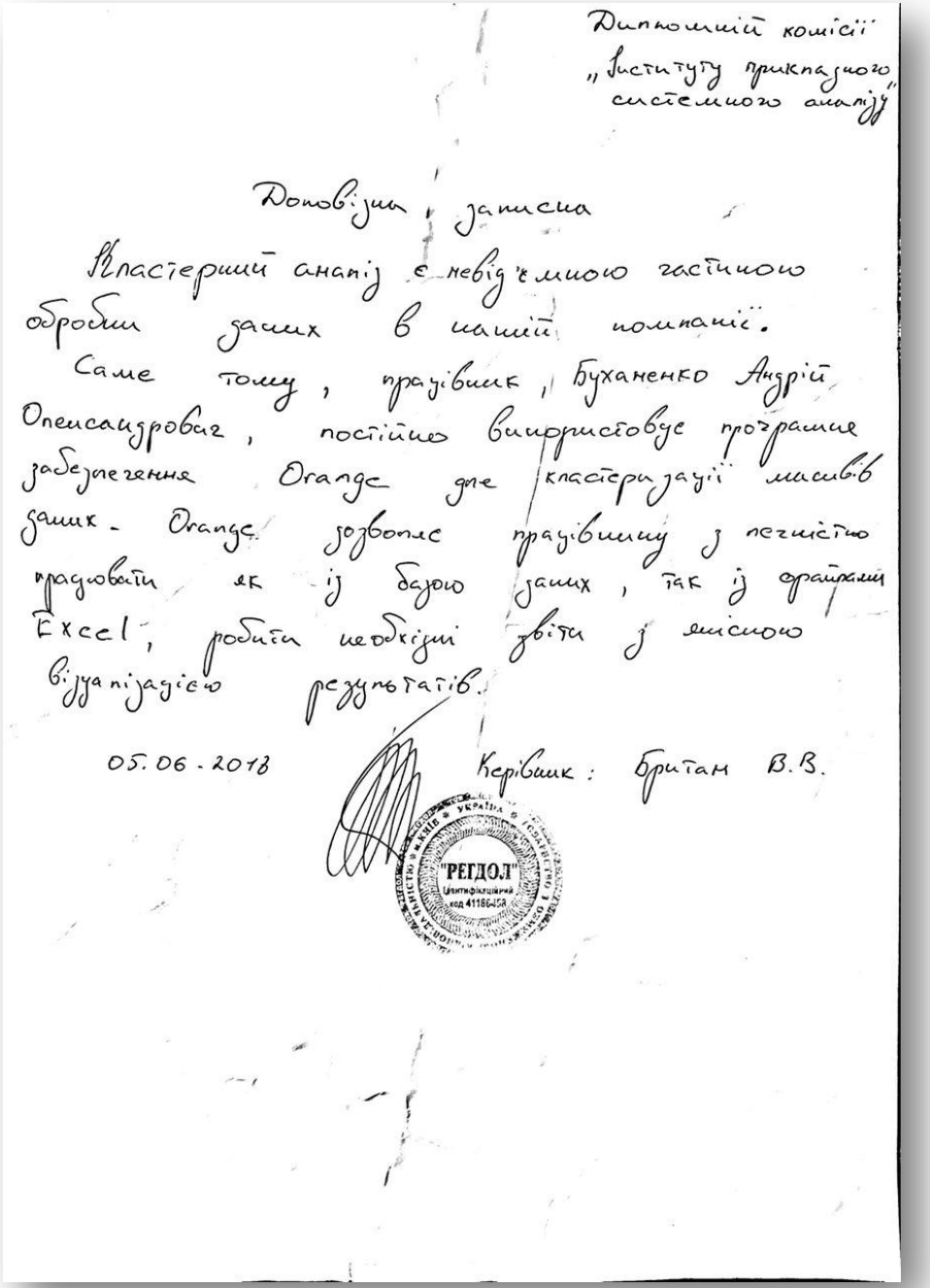

k-Means
Number of Clusters
Fixed: 15
From 2 to 20
Initialization
Random initialization
Re-runs: 10
Maximum iterations: 300
Apply Automatically
Silhouette Scores
2 0.499
3 0.508
4 0.439
5 0.435
6 0.406
7 0.332
8 0.308
9 0.281


Scatter Plot (1)
Axis Data
Axis x: Cluster
Axis y: Car
Find Informative Projections
Jittering: 10 %
Jitter numeric values
Points
Color: Car
Shape: (Same shape)
Size: (Same size)
Label: (No labels)
Symbol size:
Opacity:
Plot Properties
Show legend
Show gridlines
Show all data on mouse hover
Show class density
Show regression line
Label only selected points
Zoom/Select
Send Automatically
Volvo
VW
Toyota
Saab
Porsche
Pontiac
Olds
Nissan
Mitsub.
Mersedes
Mazda
Isuzu
Honda
Ford
Eagle
Dodge
Corvette
Chrysler
Buick
BMW
Audi
Acura
Car
C1
C2
C3
Cluster
Acura
Audi
BMW
Buick
Corvette
Chrysler
Dodge
Eagle
Ford
Honda
Other

«Hirarchical Clustering»



Акт впровадження



Дипломній комісії
„Інституту прикладного системного аналізу"

Довідка про застосування

Кластерний аналіз є невід'ємною частиною обробки даних в нашій компанії.

Саме тому, працівник, Буханенко Андрій Олександрович, постійно використовує програмне забезпечення Orange для кластеризації масивів даних. Orange дозволяє працівнику з легкістю працювати як із базою даних, так і форматами Excel, робити необхідні звіти з високою відтворюваністю результатів.

05.06.2018 Керівник: Британ В.В.

"РЕГДОЛ"
ідентифікаційний код 41186458

Методичне забезпечення

РОЗДІЛ 1	ШВИДКИЙ СТАРТ
РОЗДІЛ 2	ДЕРЕВА РІШЕНЬ
РОЗДІЛ 3	МЕТОД ОПОРНИХ ВЕКТОРІВ
РОЗДІЛ 4	ШТУЧНІ НЕЙРОННІ МЕРЕЖІ
РОЗДІЛ 5	РЕГРЕСІЙНИЙ АНАЛІЗ
РОЗДІЛ 6	КЛАСТЕРНИЙ АНАЛІЗ

Висновки

- *Розглянуто основні методи кластеризації даних, розроблена їхня порівняльна характеристика.*
- *Опрацьовані доступні методи кластеризації в програмному забезпеченні Orange.*
- *Розроблене методичне забезпечення для програмного забезпечення Orange.*

Дякую за увагу!