Разработка программного обеспечения, ориентированного на пользователя, для проведения кластер-анализа по критерию наименьших квадратов

Выполнил:

Eремейкин Пётр Александрович студент группы мНоД16-ТМСС eremeykin@gmail.com Руководитель: Миркин Борис Григорьевич д.т.н., профессор

НИУ ВШЭ Июнь 2018

Постановка задачи кластеризации

Пусть имеется N объектов и у каждого объекта определены значения V признаков. Множество всех объектов Y можно представить в виде таблицы:

$$Y = \begin{pmatrix} y_1 \\ \cdots \\ y_N \end{pmatrix} = \begin{pmatrix} y_{11} & \cdots & y_{1V} \\ \cdots & \cdots & \cdots \\ y_{N1} & \cdots & y_{NV} \end{pmatrix}$$

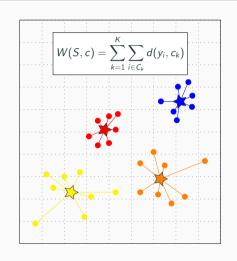
Требуется получить разбиение $S=\{C_1,\ldots,C_K\}$, состоящее из K кластеров, которые не пересекаются и покрывают всё множество объектов Y. Чёткой формулировки относительно того, что должно быть включено в кластеры, не существует. Общая идея состоит в том, чтобы сходные объекты были включены в один кластер, а несходные не принадлежали одному кластеру.

Традиционное решение (k-means)

Самый популярный алгоритм кластеризации — k-means. Он основан на поочерёдной минимизации квадратичного критерия по двум группам переменных: центрам кластеров и принадлежности объектов кластерам.

Недостатки метода:

- требует задания числа кластеров
- сильно зависит от инициализации
- плохо работает для зашумлённых данных



Предлагаемый состав программы INDACT (INtelligent DAta Clustering Toolkit)

```
• ik-means

• dePDDP

• BiKM-R

• A-Ward

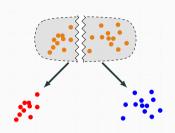
• A-Ward_{p\beta}

Алгоритмы, разработаны
```

Миркиным Б.Г.

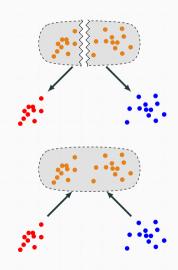
Предлагаемый состав программы INDACT (INtelligent DAta Clustering Toolkit)

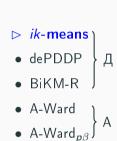
 \bullet ik-means \bullet dePDDP \bullet BiKM-R \bullet A-Ward \bullet A-Ward \bullet A-Ward

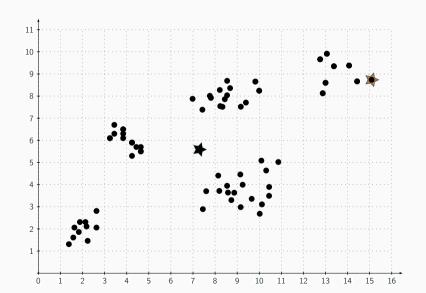


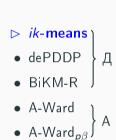
Предлагаемый состав программы INDACT (INtelligent DAta Clustering Toolkit)

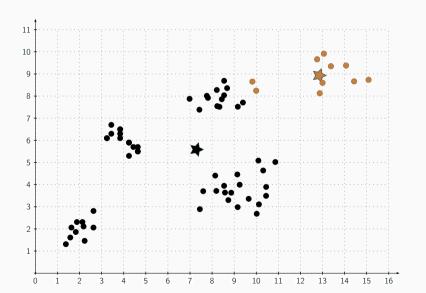
ik-means
 dePDDP
 BiKM-R
 A-Ward
 A-Ward_{pβ} **Агломеративные**

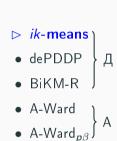


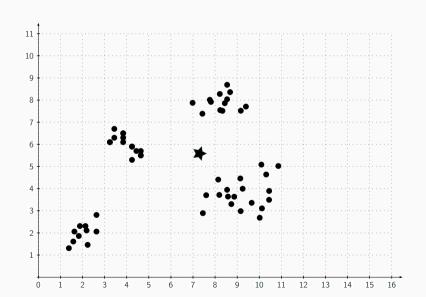


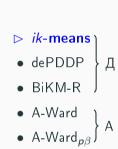


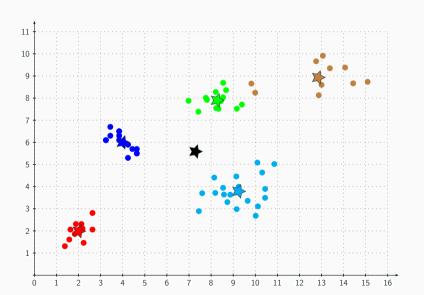


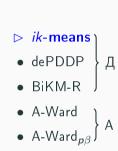


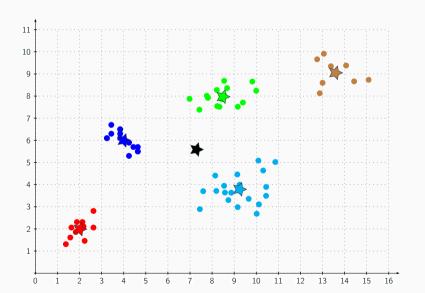










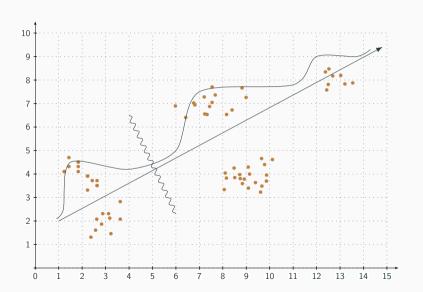


Предлагаемый состав программы INDACT: dePDDP

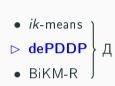


Алгоритмы разработаны Миркиным Б.Г.

• A-Ward_{$p\beta$}

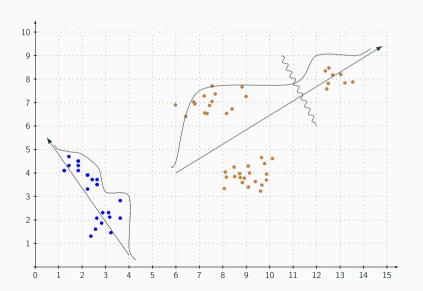


Предлагаемый состав программы INDACT: dePDDP

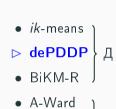


A-Ward_{pβ}
 Алгоритмы разработаны Миркиным Б.Г.

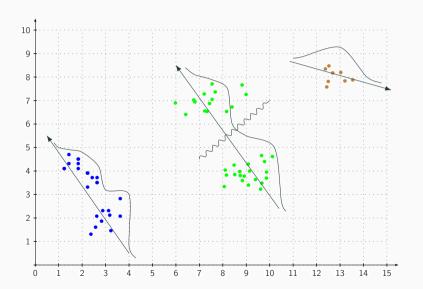
A-Ward



Предлагаемый состав программы INDACT: dePDDP



Алгоритмы разработаны Миркиным Б.Г.



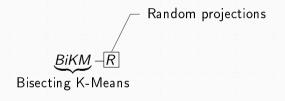
Предлагаемый состав программы INDACT: BiKM-R

- *ik*-means dePDDP
- **BiKM-R**
- A-Ward A

$$\underbrace{BiKM}_{\text{Bisecting K-Means}} - R$$

Предлагаемый состав программы INDACT: BiKM-R

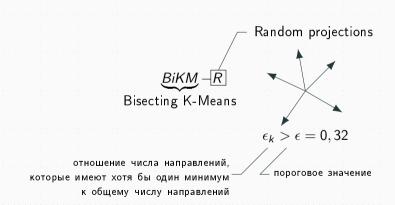




Предлагаемый состав программы INDACT: BiKM-R

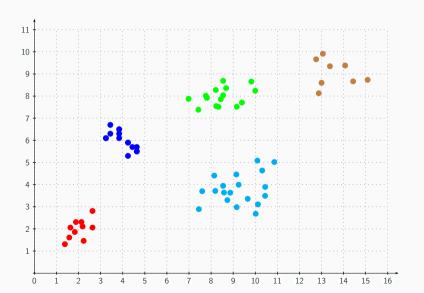


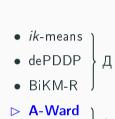
• A-Ward A-Ward A



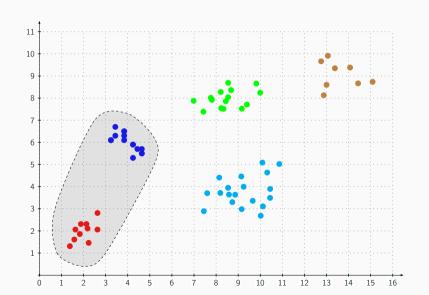


Алгоритмы разработаны Миркиным Б.Г.



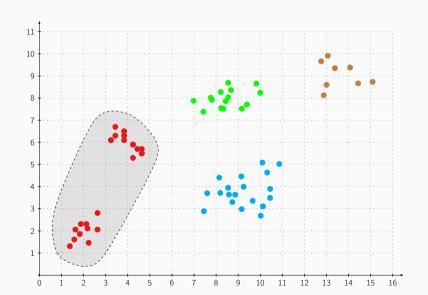


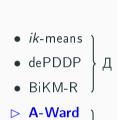
Алгоритмы разработаны Миркиным Б.Г.



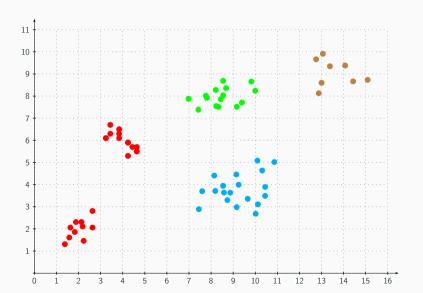


Алгоритмы разработаны Миркиным Б.Г.



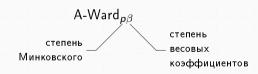


Алгоритмы разработаны Миркиным Б.Г.



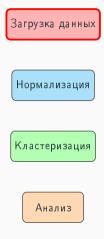
Предлагаемый состав программы INDACT: A-Ward_{рβ}

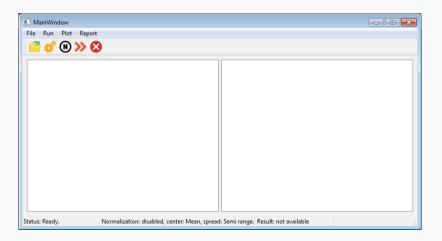




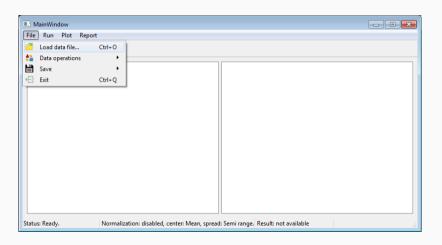
Пользовательское описание программы: стадии работы



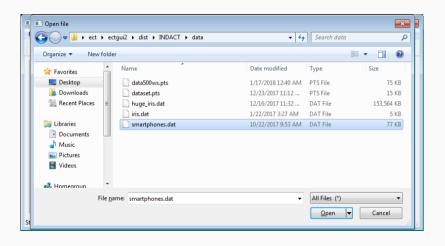


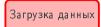






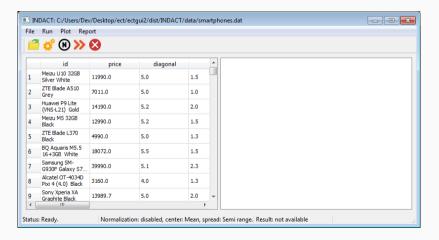






Нормализация

Кластеризация

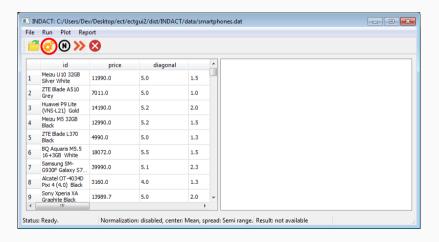


Пользовательское описание программы: нормализация

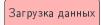
Загрузка данных

Нормализация

Кластеризация

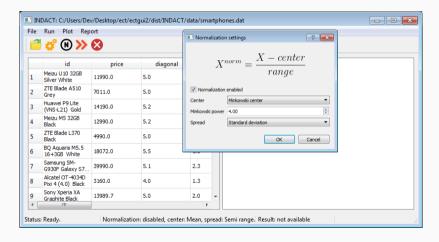


Пользовательское описание программы: нормализация



Нормализация

Кластеризация

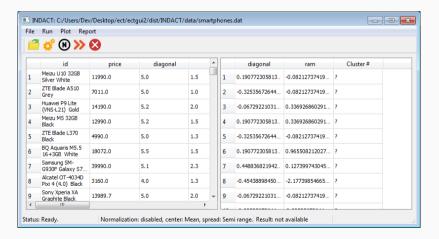


Пользовательское описание программы: нормализация

Загрузка данных

Нормализация

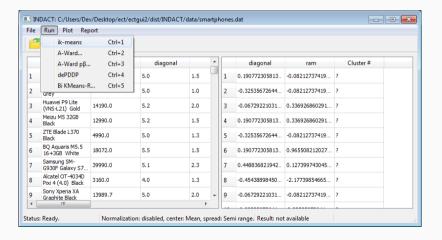
Кластеризация

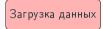


Загрузка данных

Нормализация

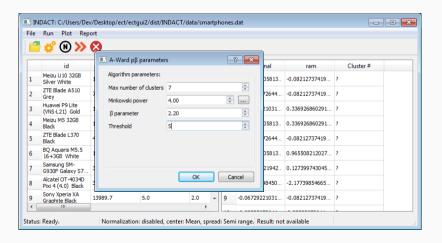
Кластеризация





Нормализация

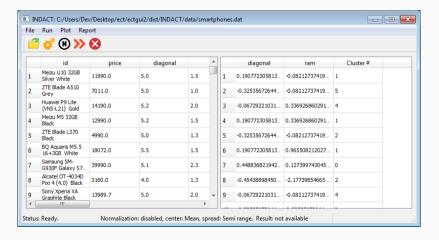
Кластеризация



Загрузка данных

Нормализация

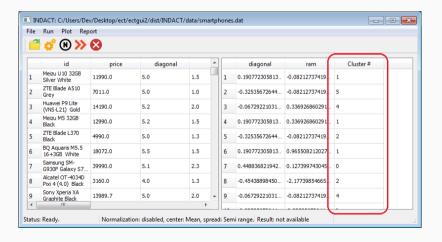
Кластеризация



Загрузка данных

Нормализация

Кластеризация

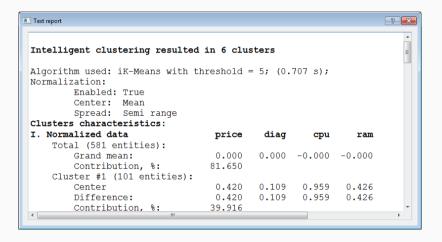


Пользовательское описание программы: анализ

Загрузка данных

Нормализация

Кластеризация

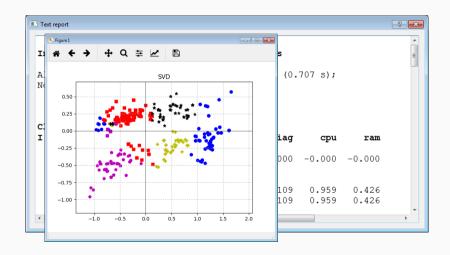


Пользовательское описание программы: анализ

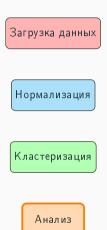
Загрузка данных

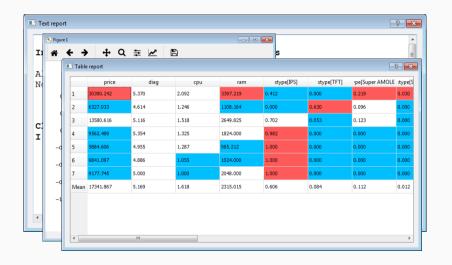
Нормализация

Кластеризация

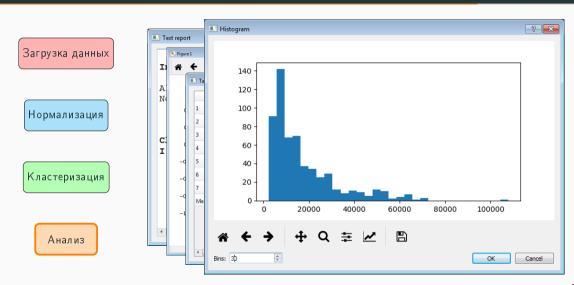


Пользовательское описание программы: анализ





Пользовательское описание программы: анализ



Внутренняя организация: основные сведения

• Язык: Python 3

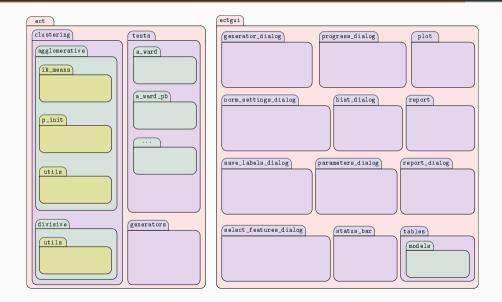
• **GUI**: PyQt

• Библиотеки: Pandas, NumPy, SciPy, scikit-learn, Matplotlib

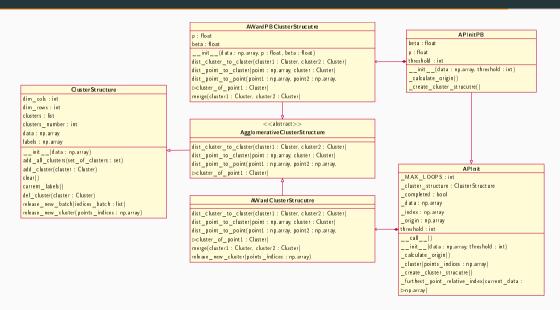
• Тестирование: pytest

• Дистрибуция: pyinstaller

Внутренняя организация: структура пакетов



Внутренняя организация: диаграмма классов



Демонстрационный пример: описание исходных данных

Размерность: 386 × 4

• Предметная область: модели смартфонов

• Источник: www.ozon.ru/context/partner xml

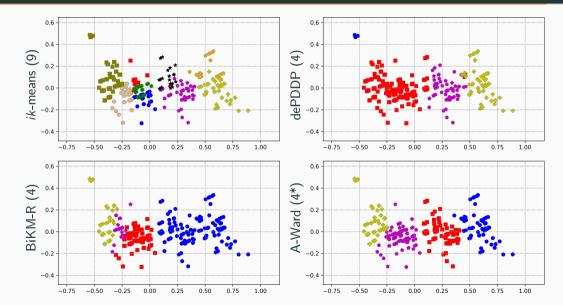
• Признаки:

#	Название	Описание	Единица	
		Описание	измерения	
1	price	Цена данной модели смартфона в IV квартале 2017 года	руб.	
2	diag	Размер диагонали экрана	дюйм	
3	сри	Частота центрального процессора (ЦП)	ГГц	
4	ram	Объем оперативной памяти	M6	

Демонстрационный пример: фрагмент данных

```
name.
                                            price, diag. cpu, ram
Meizu U10 32GB Silver White,
                                            11990.00, 5.0, 1.50, 3072
ZTE Blade A510 Grey,
                                            7011.00, 5.0, 1.00, 1024
Huawei P9 Lite (VNS-L21) Gold,
                                            14190.00, 5.2, 2.00, 2048
Meizu M5 32GB Black,
                                            12990.00, 5.2, 1.50, 3072
                                            4990.00, 5.0, 1.30, 1024
ZTE Blade L370 Black.
BQ Aquaris M5.5 16+3GB White,
                                            18072.00, 5.5, 1.50, 3072
Samsung SM-G930F Galaxy S7 (32GB) Silver, 39990.00, 5.1, 2.30, 4096
                                    3160.00, 4.0, 1.30, 512
Alcatel OT-4034D Pixi 4 (4.0) Black,
Sony Xperia XA Graphite Black,
                                            13989.70, 5.0, 2.00, 2048
ZTE Blade L5 Plus Black.
                                            6790.00, 5.0, 1.30, 1024
Meizu Pro 6 64GB Rose Gold.
                                            25990.00, 5.2, 2.50, 4096
BQ Aquaris X5 Plus Black,
                                            17290.00, 5.0, 1.80, 2048
Sony Xperia X Compact Mist Blue,
                                            24989.90, 4.6, 1.80, 3072
                                            14590.00, 5.2, 1.30, 2048
ZTE Blade V7 Rose.
Lenovo Vibe Shot (Z90A40) Red (PA1K0161RU) 16890.00, 5.0, 1.70, 3072
HTC 10 Lifestvle Topaz Gold.
                                            27990.00, 5.2, 1.80, 3072
```

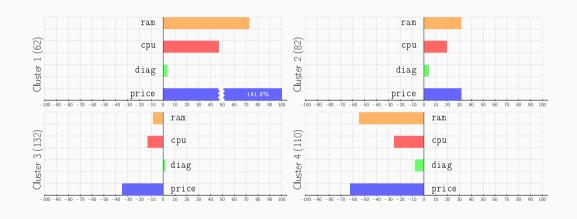
Демонстрационный пример: SVD представления разбиений



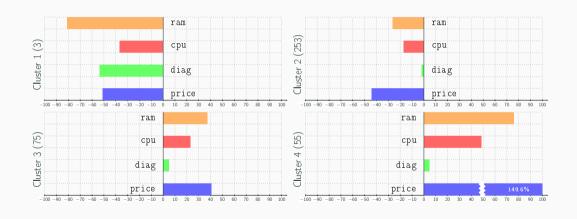
Демонстрационный пример: попарный индекс ARI

Число кластеров	Алгоритм	<i>ik</i> -means	dePDDP	BiKM-R	A-Ward
9	<i>ik</i> -means	1	0.28	0.36	0.67
4	dePDDP	-	1	0.25	0.46
4	BiKM-R	-	-	1	0.63
4*	A-Ward	-	-	-	1

Демонстрационный пример: интерпретация результатов – A-Ward



Демонстрационный пример: интерпретация результатов – dePDDP



Выводы

современных алгоритмов

• Реализована система, включающая 5

- При разработке учтены тенденции в области проектирования ПО
- INDACT позволит облегчить применение разработанных алгоритмов

Спасибо за внимание!