

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

Выполнил:

Еремейкин Пётр Александрович
студент группы мНоД16-ТМСС
eremeykin@gmail.com

Руководитель:

Миркин Борис Григорьевич
д.т.н. профессор

Постановка задачи кластеризации

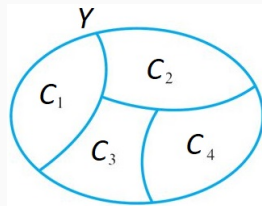
Дано: Y — множество из N объектов, характеризующихся V признаками

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

Найти: разбиение $S = \{C_1, \dots, C_K\}$, из K кластеров,

$$\bigcup_{k=1}^K C_k = Y$$

$$C_{k1} \cap C_{k2} = \emptyset, k1 \neq k2$$



INDACT: INtelligent DAta Clustering Toolkit

Система для кластерного анализа данных
(~ 5000 объектов, ~ 30 признаков)

- **Язык:** Python 3
- **Пользовательский интерфейс:** оконный графический (PyQt, Matplotlib)
- **Состав:** 5 новейших алгоритмов кластеризации
- **Библиотеки:** Pandas, NumPy, SciPy, scikit-learn
- **Тестирование:** pytest
- **Дистрибуция:** pyinstaller

- Графический интерфейс
- Библиотека алгоритмов:
 - * нормализация
 - * кластер-анализ
 - * подготовка отчётов
 - * проведение вычислительных экспериментов

Пользовательский графический интерфейс

Загрузка данных

Нормализация

Кластеризация

Подготовка отчёта

The screenshot displays the INDACT application window with the file path `C:/Users/Dev/Desktop/ect/ectgui2/dist/INDACT/data/smartphones.dat`. The main window contains a menu bar (File, Run, Plot, Report) and a toolbar with icons for file operations, settings, and execution. Below the toolbar is a table of smartphone data.

	id	price	diagonal	
1	Meizu U10 32GB Silver White	11990.0	5.0	
2	ZTE Blade A510 Grey	7011.0	5.0	
3	Huawei P9 Lite (VNS-L21) Gold	14190.0	5.2	
4	Meizu M5 32GB Black	12990.0	5.2	
5	ZTE Blade L370 Black	4990.0	5.0	
6	BQ Aquaris M5.5 16+3GB White	18072.0	5.5	
7	Samsung SM-G930F Galaxy S7...	39990.0	5.1	2.3
8	Alcatel OT-4034D Pixi 4 (4.0) Black	3160.0	4.0	1.3
9	Sony Xperia XA Graphite Black	13989.7	5.0	2.0

Overlaid on the main window is a "Normalization settings" dialog box. It features the formula
$$X^{norm} = \frac{X - center}{range}$$
 and the following options:

- ☒ Normalization enabled
- Center: Minkowski center (dropdown menu)
- Minkowski power: 4.00 (spin box)
- Spread: Standard deviation (dropdown menu)

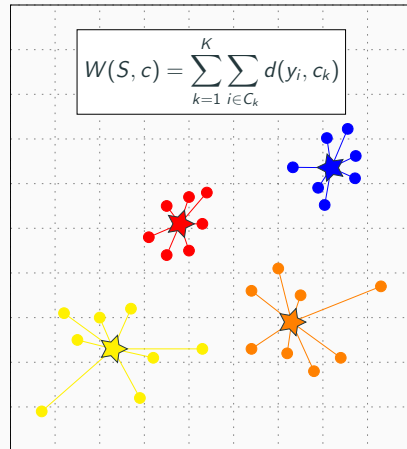
Buttons for "OK" and "Cancel" are at the bottom of the dialog. The status bar at the bottom of the main window reads: "Status: Ready. Normalization: disabled, center: Mean, spread: Semi range. Result: not available".

k -means

Поочерёдная минимизация квадратичного критерия по двум группам переменных: центрам кластеров и принадлежности объектов кластерам.

Недостатки:

- число кластеров
- инициализация
- неточности в данных
(лишние признаки, объекты и пр.)

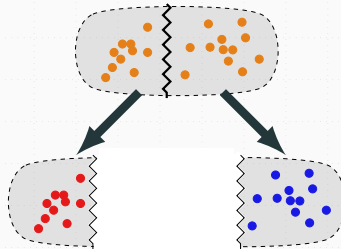


- *ik*-means
 - dePDDP
 - BiKM-R
- } Дивизивные
- A-Ward
 - A-Ward _{$p\beta$}
- } Агломеративные

Алгоритмы разработаны
Миркиным et al.

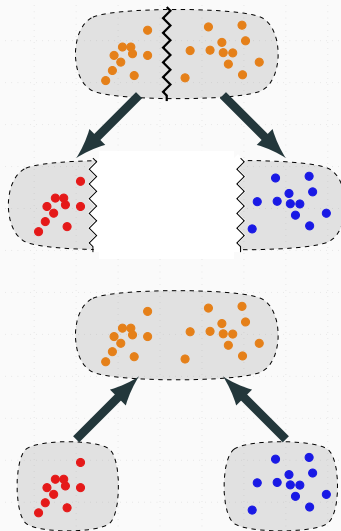
- *ik*-means
 - dePDDP
 - BiKM-R
- } Дивизивные
- A-Ward
 - A-Ward _{$p\beta$}
- } Агломеративные

Алгоритмы разработаны
Миркиным et al.



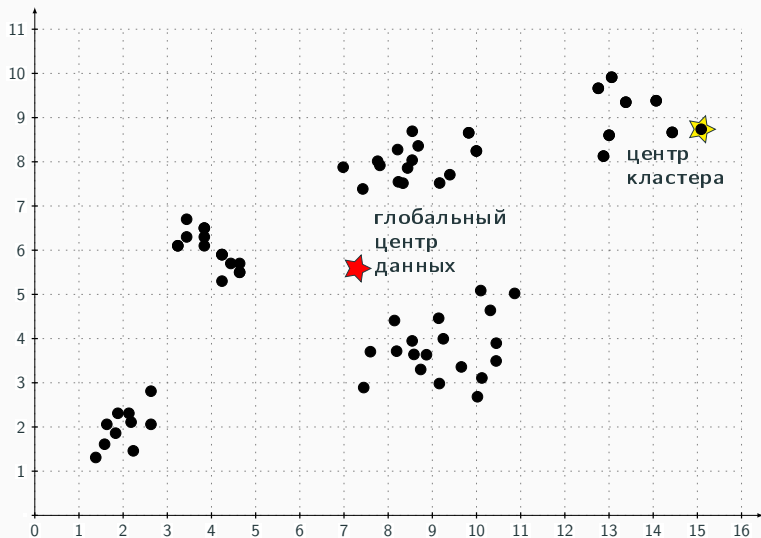
- *ik*-means
 - dePDDP
 - BiKM-R
- } Дивизивные
- A-Ward
 - A-Ward _{$p\beta$}
- } Агломеративные

Алгоритмы разработаны
Миркиным et al.



- ▷ *ik-means*
- dePDDP } Д
 - BiKM-R }
 - A-Ward } А
 - A-Ward _{$p\beta$} }

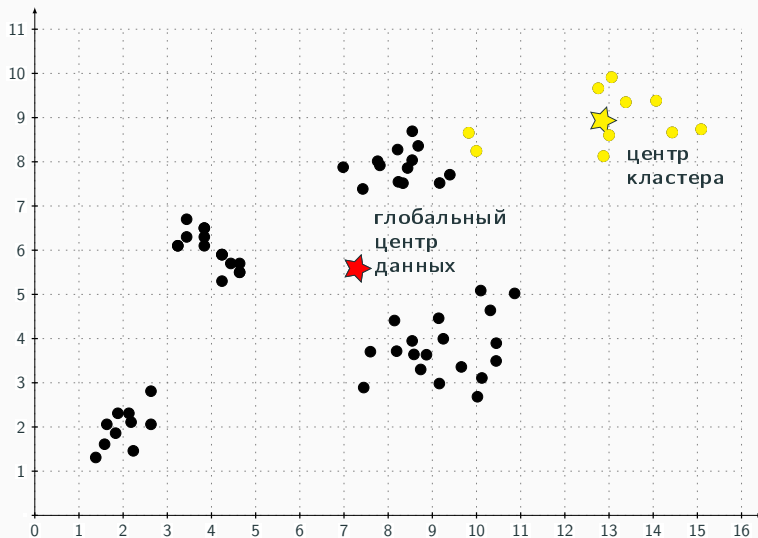
Алгоритмы разработаны
Миркиным et al.



▷ *ik-means*

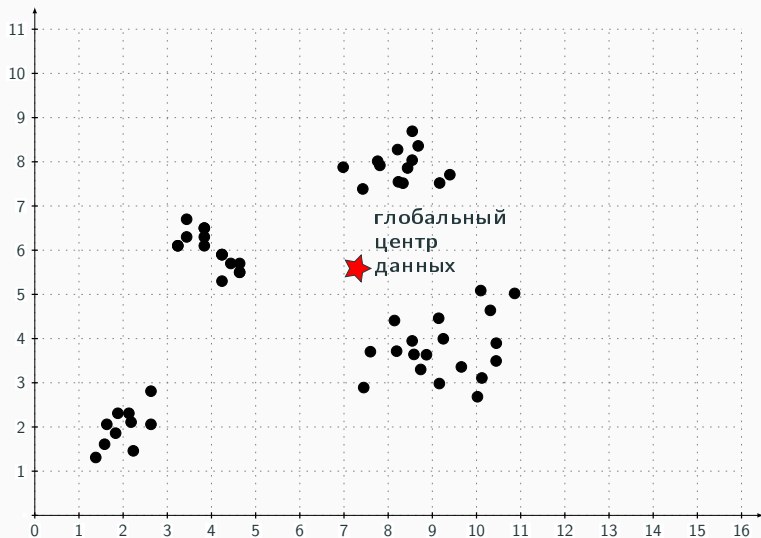
- dePDDP } Д
- BiKM-R }
- A-Ward } А
- A-Ward_{pβ} }

Алгоритмы разработаны
Миркиным et al.



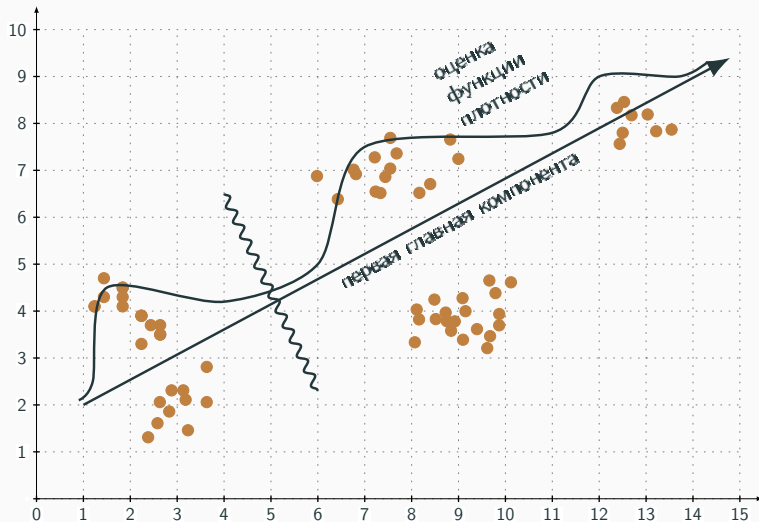
- ▷ *ik-means*
- dePDDP } Д
 - BiKM-R }
 - A-Ward } А
 - A-Ward _{$p\beta$} }

Алгоритмы разработаны
Миркиным et al.



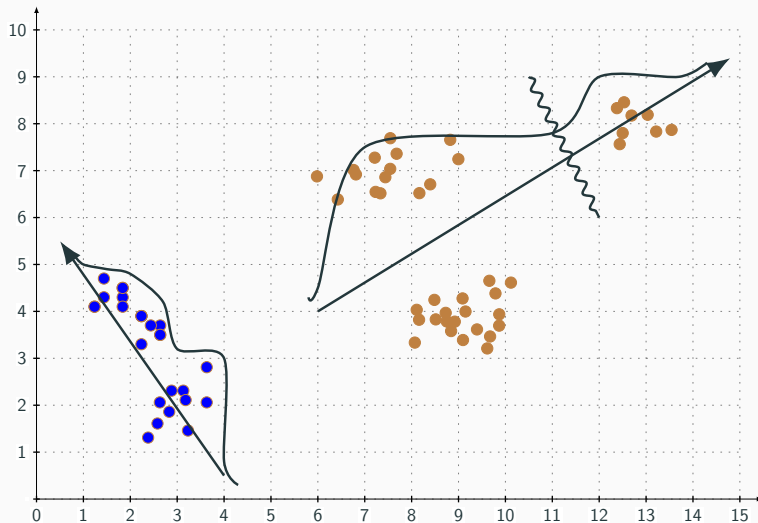
- *ik-means*
 - ▶ **dePDDP**
 - BiKM-R
- } Д
- A-Ward
 - $A\text{-Ward}_{p\beta}$
- } А

Алгоритмы разработаны
Миркиным et al.



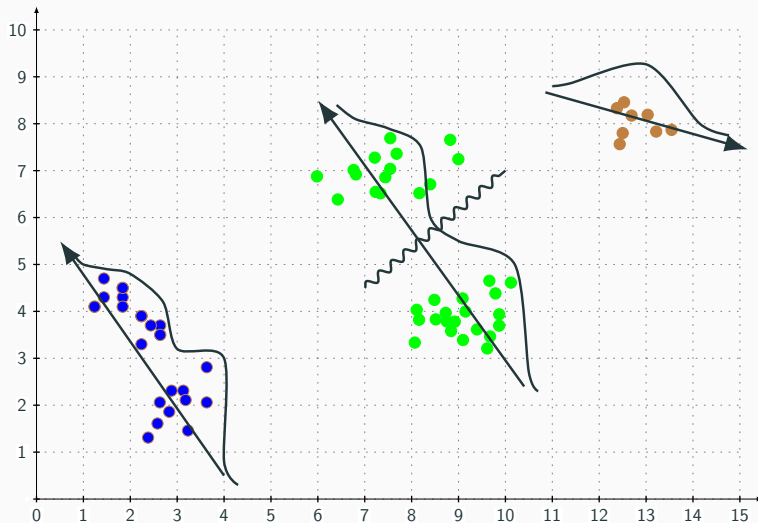
- *ik*-means
 - ▷ **dePDDP**
 - BiKM-R
- } Д
- A-Ward
 - A-Ward _{$p\beta$}
- } А

Алгоритмы разработаны
Миркиным et al.



- *ik*-means
 - ▷ **dePDDP**
 - BiKM-R
- } Д
- A-Ward
 - A-Ward _{$p\beta$}
- } А

Алгоритмы разработаны
Миркиным et al.



- *ik*-means
 - dePDDP
 - ▷ **BiKM-R**
- } Д
- A-Ward
 - A-Ward _{$p\beta$}
- } А

Алгоритмы разработаны
Миркиным et al.

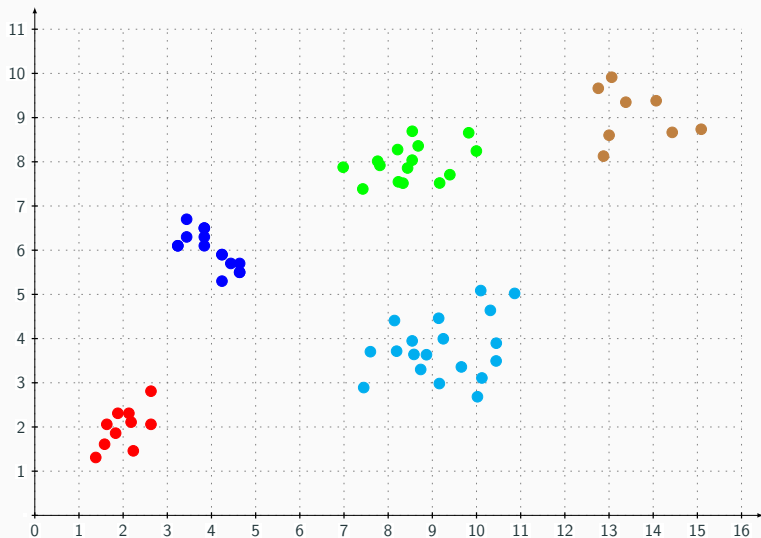
Random projections

BiKM - R

Bisecting K-Means

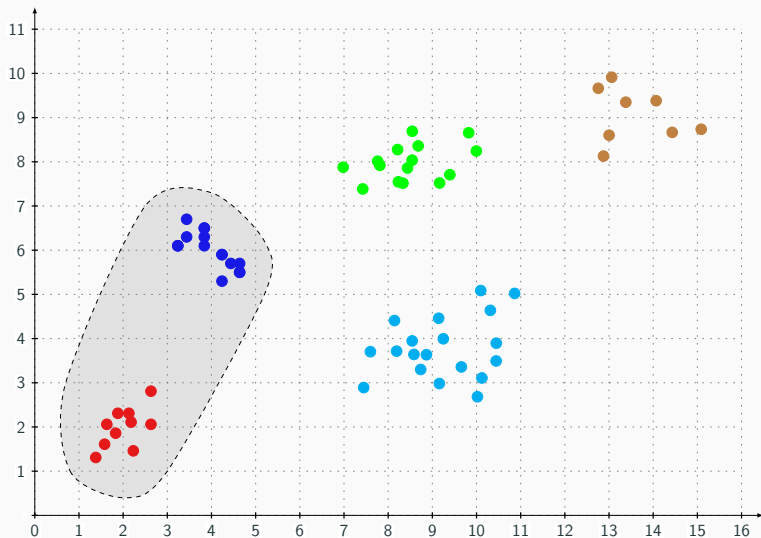
- *ik*-means
 - dePDDP
 - BiKM-R
- } Д
- ▷ **A-Ward**
 - A-Ward _{$p\beta$}
- } А

Алгоритмы разработаны
Миркиным et al.



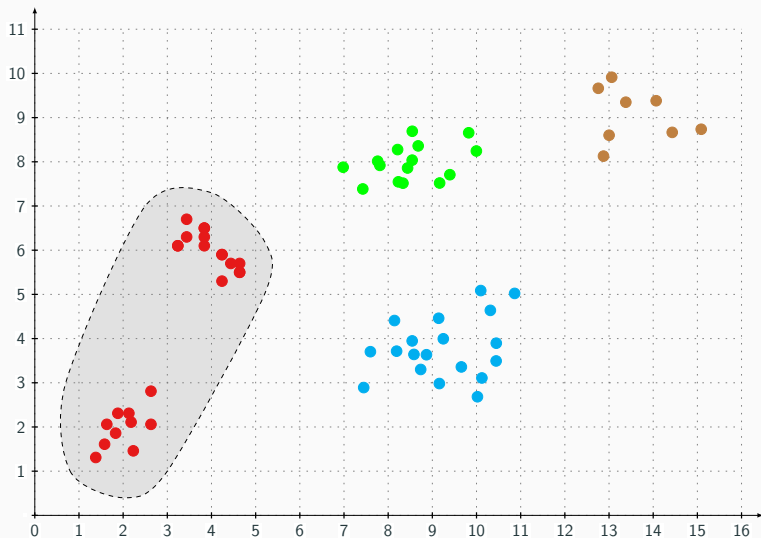
- *ik*-means
 - dePDDP
 - BiKM-R
 - ▷ **A-Ward**
 - A-Ward _{$p\beta$}
- } Д
- } А

Алгоритмы разработаны
Миркиным et al.



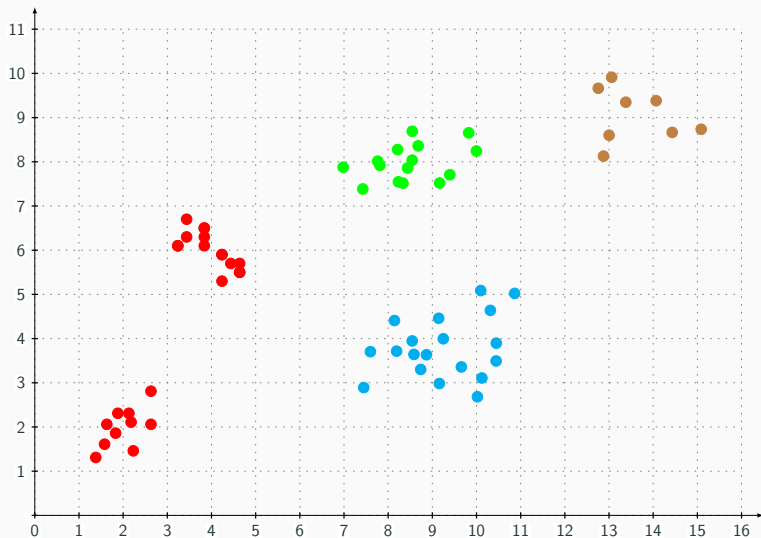
- *ik*-means
 - dePDDP
 - BiKM-R
- } Д
- ▷ **A-Ward**
 - A-Ward _{$p\beta$}
- } А

Алгоритмы разработаны
Миркиным et al.



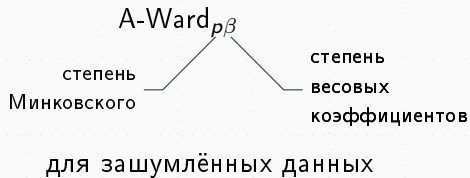
- *ik*-means
 - dePDDP
 - BiKM-R
- } Д
- ▷ **A-Ward**
 - A-Ward _{$p\beta$}
- } А

Алгоритмы разработаны
Миркиным et al.



- ik -means
 - dePDDP
 - BiKM-R
 - A-Ward
- ▷ **A-Ward _{$p\beta$}**

Алгоритмы разработаны
Миркиным et al.



Демонстрационный пример

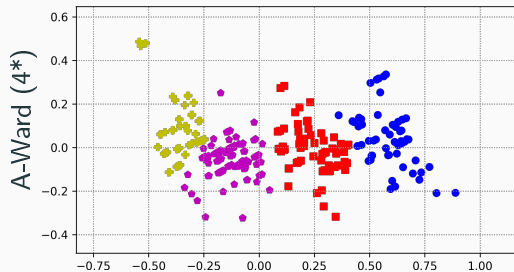
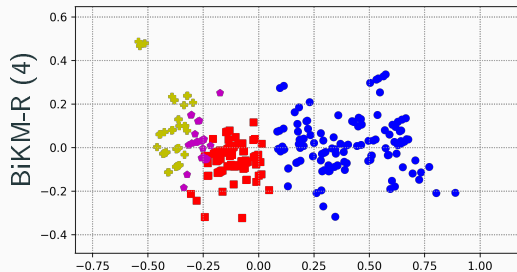
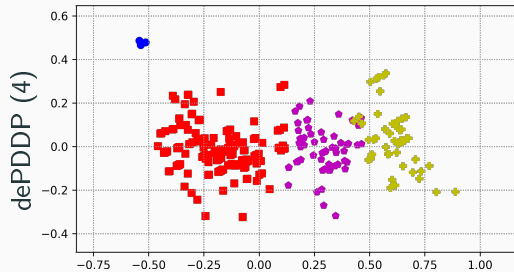
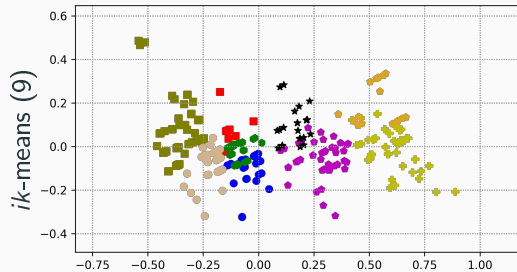
- **Размерность:** 386×4
- **Предметная область:** модели смартфонов
- **Источник:** www.ozon.ru/context/partner_xml
- **Признаки:**

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

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
ZTE Blade L370 Black ,	4990.00 ,	5.0 ,	1.30 ,	1024
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
Alcatel OT-4034D Pixi 4 (4.0) Black ,	3160.00 ,	4.0 ,	1.30 ,	512
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
ZTE Blade V7 Rose ,	14590.00 ,	5.2 ,	1.30 ,	2048
Lenovo Vibe Shot (Z90A40) Red (PA1K0161RU)	16890.00 ,	5.0 ,	1.70 ,	3072
HTC 10 Lifestyle Topaz Gold ,	27990.00 ,	5.2 ,	1.80 ,	3072

Демонстрационный пример: SVD представления разбиений

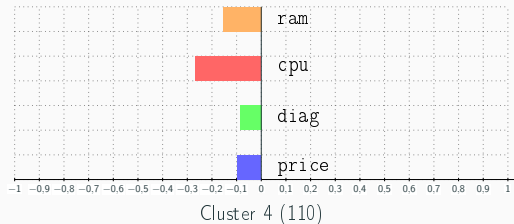
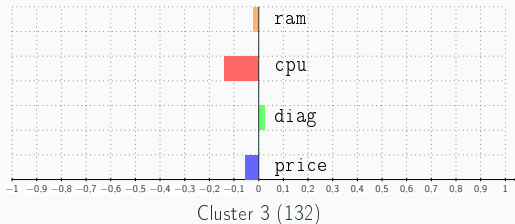
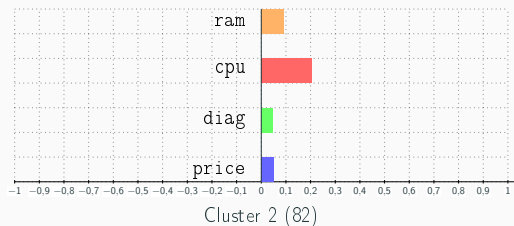
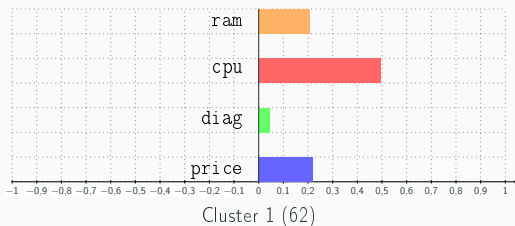


Полученное число кластеров

Число кластеров	Алгоритм
9	<i>ik</i> -means
4	dePDDP
4	BiKM-R
4*	A-Ward

A-Ward (наиболее согласованный)

Отклонения от среднего в каждом кластере в нормализованной шкале (-1...1)



Выводы

- Реализована система, включающая 5 современных алгоритмов
- Для системы INDACT разработана инструкция пользователя
- При разработке учтены тенденции в области проектирования ПО
- INDACT позволит облегчить применение разработанных алгоритмов

Спасибо за внимание!