

28/04/2018 v3.1

НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ
“ВЫСШАЯ ШКОЛА ЭКОНОМИКИ”

“СИСТЕМА ИНТЕЛЛЕКТУАЛЬНОЙ
КЛАСТЕРИЗАЦИИ ДАННЫХ”
(INTELLIGENT DATA CLUSTERING TOOLKIT, INDACT)

ИНСТРУКЦИЯ ПОЛЬЗОВАТЕЛЯ

Разработчик:
Еремейкин П.А.
студент группы
мНод16_ТМСС

Руководитель:
профессор
Миркин Б.Г.

Москва 2018

Аннотация

Система интеллектуальной кластеризации данных INDACT представляет собой программный комплекс, предназначенный для проведения кластер-анализа с применением интеллектуальных подходов. Задача кластер-анализа состоит в разделении таблицы объектов в множества (кластеры) таким образом, чтобы сходные объекты попали в один и тот же кластер, а несходные — в разные. Широко известен традиционный метод k -средних. Однако, этот метод обладает существенным недостатком: для его применения необходимо знать число кластеров, на которые будут разбиты данные. В таких случаях задачу позволяют решить интеллектуальные методы, которые позволяют автоматически определить число кластеров. Программная система INDACT обладает всем необходимым функционалом и включает в свой инструментарий множество методов, необходимых для эффективного решения сложных задач кластер анализа.

Содержание

1	Введение	4
1.1	Область применения	4
1.2	Возможности системы INDAST	4
1.3	Уровень подготовки пользователя	4
1.4	Исходный код	5
2	Назначение	6
3	Подготовка к работе	7
4	Организация работы	8
4.1	Этапы	8
4.2	Требования к файлу исходных данных	8
4.3	Обучающий файл	9
4.4	Нормализация данных	9
4.5	Просмотр результатов кластеризации	10
4.6	Элементы пользовательского интерфейса	11
4.6.1	Главное окно	11
4.6.2	Контекстное меню	13
4.6.3	Диалог нормализации	14
4.6.4	Окно графического вывода	15
4.6.5	Окно генерации данных	16
4.6.6	Окно запуска кластеризации	17
4.6.7	Окно таблицы результатов	19
5	Описание операций	21
5.1	Запуск программы	21
5.2	Загрузка исходных данных	22
5.3	Нормализация	23
5.3.1	Установка параметров нормализации	23
5.3.2	Нормализация одного признака	25
5.3.3	Нормализация нескольких признаков сразу	27
5.4	Отбор признаков	29
5.4.1	Удаление одного признака	29
5.4.2	Удаление нескольких признаков сразу	30
5.5	Визуализация	32
5.5.1	Построение гистограммы признака	32
5.5.2	Построение поля рассеяния (scatter plot)	34

5.5.3	Построение SVD диаграммы	37
5.6	Генерация синтетических данных	38
5.7	Запуск кластеризации	40
5.8	Генерация отчёта	41
5.8.1	Текстовый отчёт	41
5.8.2	Табличный отчёт	43
5.9	Выход из программы	44
6	Алгоритмы кластеризации (краткое описание)	45
6.1	Алгоритм A-Ward	45
6.2	Алгоритм A-Ward _{$p\beta$}	45
6.3	Алгоритм BiKM-R	47
6.4	Алгоритм dePDDP	47
7	Демонстрация работы программы	50
7.1	Нормализация	50
7.1.1	Нормализация с центрированием по среднему и масштабировани- ем по полуразмаху	50
7.1.2	Нормализация с центрированием по Минковскому и масштабиро- ванием по стандартному отклонению	54
7.2	Кластеризация	55
7.2.1	Кластеризация с автоматическим выбором числа кластеров	55
7.2.2	Кластеризация с заданным числом кластеров	62
	Аббревиатуры	64
	Словарь терминов	65
	Список литературы	66

1 Введение

1.1 Область применения

Программное обеспечение INDAST применяется для проведения кластер-анализа таблиц данных с использованием интеллектуальных алгоритмов. Типичный пример данных к которым может применяться кластер-анализ — ирисы Фишера. Может ли метод кластер-анализа воспроизвести три естественных таксона (вида) — Ирис щетинистый (*Iris setosa*), Ирис виргинский (*Iris virginica*) и Ирис разноцветный (*Iris versicolor*), к которым принадлежат ирисы. Данные представляют собой таблицу 150×4 значений четырех признаков:

1. Длина чашелистика
2. Ширина чашелистика
3. Длина лепестка
4. Ширина лепестка

на 150 экземплярах цветов, собранных ботаником Андерсоном, утверждавшим, что каждый таксон представлен 50 экземплярами. Кластер-анализ применяется во многих областях, включая компьютерное зрение, маркетинг, биоинформатику и медицину[1].

1.2 Возможности системы INDAST

Функционал:

- просмотр таблиц данных
- нормализация данных
- проведение кластер-анализа с использованием арсенала эффективных методов
- визуализация результатов
- генерация отчёта
- проведение вычислительных экспериментов с генерацией синтетических данных

1.3 Уровень подготовки пользователя

Для работы с программой от пользователя требуется знание основ работы с графическим интерфейсом современных операционных систем (ОС).

1.4 Исходный код

Программа обладает открытым исходным кодом. Исходный код программы можно получить из github репозитория по следующим ссылкам:

1. <https://github.com/eremeykin/ect> — репозиторий с исходным кодом библиотек кластеризации (для вызова из Python программ)
2. <https://github.com/eremeykin/ectgui2> — репозиторий с графической оболочкой, которая использует библиотеку кластеризации.

Для запуска интерфейса программы из исходных кодов потребуется скачать оба репозитория. Если необходимо только использовать реализованные алгоритмы кластеризации, вызывая их из другой Python программы, потребуется только первый репозиторий.

2 Назначение

Система интеллектуальной кластеризации INDACT предназначена для формирования множеств объектов (кластеров) таким образом, чтобы сходные объекты попадали в один и тот же кластер, а несходные — в разные кластеры [2]. Основной целью INDACT является повышение эффективности анализа данных за счет использования наиболее продвинутых в теоретическом и практическом отношении алгоритмов.

3 Подготовка к работе

Программный продукт работает в операционной системе Microsoft Windows 7 ¹ со следующими характеристиками:

- объем ОЗУ не менее 2 Гб
- объем жесткого диска не менее 40 Гб
- микропроцессор с тактовой частотой не менее 1.5 Гц
- монитор с разрешением от 1280 × 1024 точек и выше

Все программные компоненты уже включены в распространяемый каталог, установка интерпретатора Python и специальных библиотек не требуется. Необходимые dll библиотеки и другие ресурсы также находятся в каталоге бинарных файлов в полном составе.

Для подготовки системы к работе требуется скопировать каталог с бинарными файлами программы с носителя на котором распространяется программа на запоминающее устройство ПК пользователя. Каталог бинарных файлов назван **INDACT**. Для начала работы пользователь запускает на выполнение файл **INDACT.exe** из каталога бинарных файлов. При необходимости, пользователь может создать ярлык на исполняемый файл и запускать программу из любого удобного места.

В каталоге бинарных файлов пользователь может также найти каталог **data**, в котором собраны некоторые иллюстративные наборы данных. Загрузка этих файлов в программу протестирована и не вызывает ошибок, поэтому их можно использовать для понимания структуры загружаемых файлов. Подробнее о требованиях к файлам данных см. раздел [4.3 Обучающий файл](#).

Настройки программа хранит в файле **settings.ini**. При необходимости, пользователь может удалить его, чтобы сбросить все настройки или изменять вручную (как правило, такой необходимости нет).

¹ Программная система разработана на кроссплатформенном языке программирования Python и может быть запущена также на других операционных системах, при условии удовлетворения всех необходимых зависимостей. Установка библиотек для различных ОС выходит за рамки инструкции. Версия для Windows специально подготовлена для использования без необходимости установки интерпретатора или программных библиотек.

4 Организация работы

4.1 Этапы

Работа с программой INDACT строится на основе графического диалогового интерфейса. Типичный сценарий взаимодействия пользователя с программой разделяется на следующие этапы:

1. Запуск программы
2. Загрузка данных
3. Нормализация данных
4. Отбор признаков
5. Выполнение кластеризации
6. Визуализация результатов
7. Формирование текстового отчёта

После запуска программы требуется выбрать файл, содержащий данные для кластеризации. Затем производится настройка параметров нормализации (см. 4.4), отбор признаков, участвующих в кластеризации и выбор основных свойств применяемого алгоритма. После выбора необходимых параметров пользователь запускает алгоритм кластеризации. Когда выполнение кластеризации заканчивается, пользователю становятся доступны результаты работы для просмотра, анализа и сохранения.

4.2 Требования к файлу исходных данных

Источником данных для программы является текстовый файл. Следует уделить особое внимание формату файла. Ниже перечислены требования к загружаемому файлу:

1. Файл содержит записи в формате таблицы объект-признак
2. Строки таблицы соответствуют объектам
3. Столбцы таблицы соответствуют признакам
4. Разделитель строк — символ перевода строки (CR+LF для Windows)
5. Разделитель столбцов — запятая
6. Первая строка обязательно содержит перечень названий признаков

7. Названия признаков состоят только из латинских букв
8. Разделитель дробной и целой части — точка
9. Значения номинальных признаков записываются в одно слово из латинских букв. Цифры не допустимы.

Пример файла с валидной структурой приведён в разделе [4.3 Обучающий файл](#).

4.3 Обучающий файл

Демонстрация возможностей программы будет проиллюстрирована на обучающем наборе данных. Файл `smartphones.dat` с демонстрационной таблицей данных можно найти в каталоге бинарных файлов программы в директории `data`. Этот файл можно открыть с помощью текстового редактора, например стандартного блокнота Windows и при необходимости отредактировать или просто посмотреть содержимое.

Демонстрационный файл содержит 581×7 таблицу параметров смартфонов, продаваемых в магазине Ozon (<http://www.ozon.ru/>) в IV квартале 2017 года. Каждая строка таблицы соответствует определённому смартфону и показывает значения 7 параметров: `name`, `price`, `diag`, `cpu`, `ram`, `stype`, `vendor`; т.е. название смартфона, цена в рублях, диагональ экрана в дюймах, частота процессора в ГГц, объем ОЗУ в Мб, тип матрицы, вендор.

Фрагмент этого файла приведён ниже. Названия сокращены в целях наглядности.

Пример файла входных данных `smartphones.dat`

<code>name</code> ,	<code>price</code> ,	<code>diag</code> ,	<code>cpu</code> ,	<code>ram</code> ,	<code>stype</code> ,	<code>vendor</code>
Meizu U10 32GB,	11990.00,	5.0,	1.50,	3072,	IPS,	Meizu
ZTE Blade A510,	7011.00,	5.0,	1.00,	1024,	IPS,	ZTE
Huawei P9 Lite,	14190.00,	5.2,	2.00,	2048,	IPS,	Huawei
Meizu M5 32GB ,	12990.00,	5.2,	1.50,	3072,	IPS,	Meizu
ZTE Blade L370,	4990.00,	5.0,	1.30,	1024,	TFT,	ZTE
BQ Aquaris M5 ,	18072.00,	5.5,	1.50,	3072,	IPS,	BQ

4.4 Нормализация данных

Нормализация — это преобразование данных для приведения всех признаков к сопоставимым шкалам и началам отсчёта. Общая формула нормализации может быть записана следующим образом:

$$X' = \frac{X - c}{r}, \quad (1)$$

где X — исходные данные,

c — параметр, определяющий преобразование начала отсчёта,

r — параметр, определяющий преобразование масштаба шкалы.

В некоторых случаях нормализация влияет существенным образом на результат кластеризации. В программе INDACT реализованы наиболее популярные способы определения параметра преобразования начала отсчёта:

- среднее
- минимум
- медиана
- центр Минковского

Для параметра, определяющего масштаб шкалы, предусмотрены следующие способы вычисления:

- полуразмах
- стандартное отклонение
- абсолютное отклонение

В системе INDACT процедура нормализации реализована независимо от этапа кластеризации, и параметры нормализации могут быть изменены практически на любой стадии работы с системой. Как правило, нормализация данных проводится сразу после их загрузки. Этап нормализации можно пропустить, если данные уже нормированы или в этом нет необходимости.

Выполнению кластеризации предшествует выбор параметров и принципов, на которых основывается процесс поиска однородных множеств. После выбора всех необходимых параметров пользователь производит запуск алгоритма и получает результат в интерфейсе программы.

4.5 Просмотр результатов кластеризации

Просмотр результатов кластеризации может состоять в отслеживании принадлежности каждого объекта определенным кластерам или получении графического представления найденной кластерной структуры. Также система INDACT позволяет представить результат в виде кластерной таблицы или в виде текстового отчёта.

4.6 Элементы пользовательского интерфейса

4.6.1 Главное окно

Как было отмечено ранее, программа обладает графическим пользовательским интерфейсом. В данном разделе приведены основные сведения относительно элементов управления, их положения и функциях.

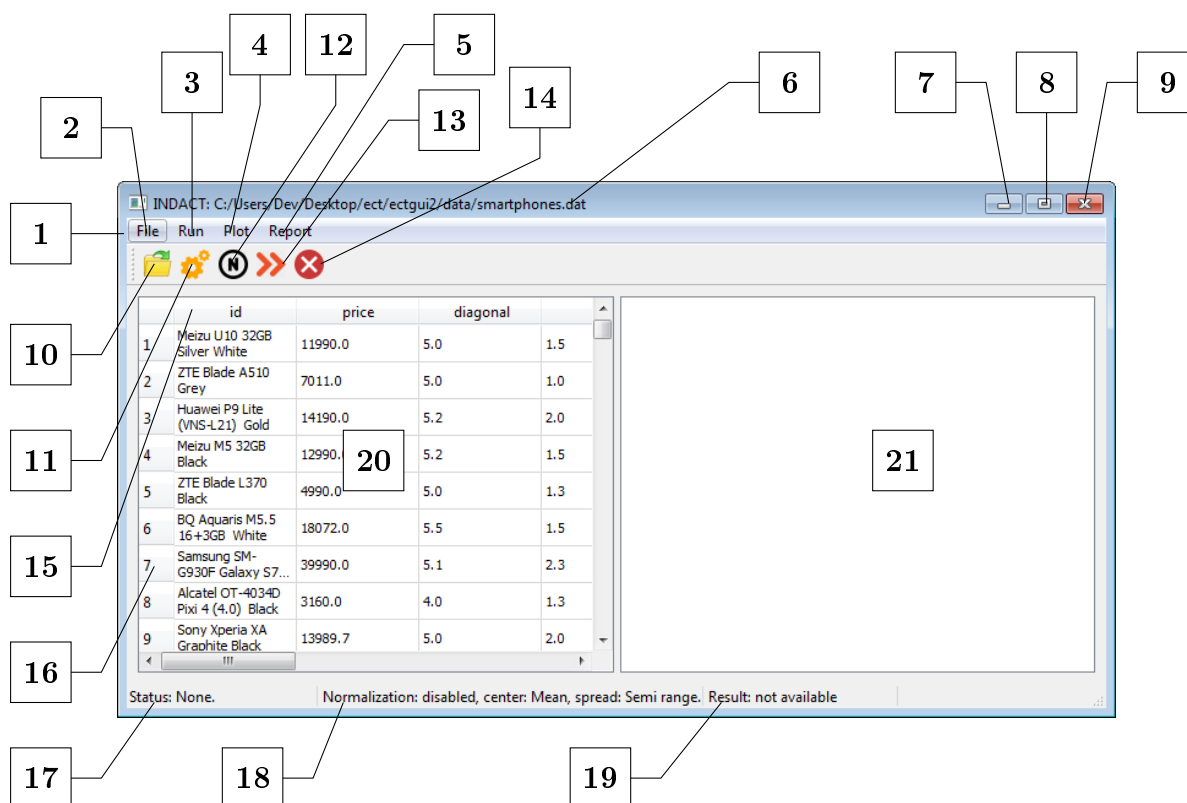


Рисунок 1 – Основные элементы пользовательского интерфейса

Цифры на рисунке означают:

1. Главное меню, элемент интерфейса, содержащий основные команды
2. Меню File, содержит пункты:
 - Load data file (Ctrl + O) — для загрузки файла данных
 - Data operations — для манипулирования данными
 - Exit (Ctrl + Q) — для выхода из программы
3. Меню Run, содержит пункты для запуска соответствующих алгоритмов:

- A-Ward (Ctrl+ 1)
 - A-Ward_{pβ} (Ctrl+ 2)
 - Bi K-Means R (Ctrl+ 3)
 - dePDDP (Ctrl+ 4)
 - IK-Means (Ctrl+ 5)
4. Меню Plot, служит для вызова команд графического отображения, содержит пункты:
- By Markers — для построения поля рассеяния (scatter plot) по отмеченным признакам
 - SVD — для построения SVD диаграммы
 - Remove markers — для удаления всех отметок признаков
5. Меню Report для формирования отчёта, содержит пункты:
- Text (Ctrl+ R)— для отображения текстового отчёта
 - Text — для отображения табличного отчёта
6. Заголовок окна, содержит путь к открытому файлу
7. Кнопка “Свернуть окно”
8. Кнопка “Развернуть окно”
9. Кнопка “Закрыть окно”
10. Иконка “Загрузить данные”, дублирует соответствующий пункт меню
11. Иконка “Настройки” вызывает диалог настроек нормализации
12. Иконка включения/выключения нормализации данных
13. Иконка нормализации нескольких признаков сразу
14. Иконка очистки нормализованных признаков
15. Названия признаков
16. Номера/названия объектов
17. Строка состояния, выводит информацию о выполняемом действии
18. Текущие параметры нормализации

19. Последний результат кластеризации
20. Панель с исходными данными
21. Панель с нормализованными данными

4.6.2 Контекстное меню

В данном разделе описаны пункты контекстного меню. Контекстное меню объединяет набор действий над определенным объектом и вызывается щелчком правой кнопки мыши на этом объекте. На рисунке 2 показано контекстное меню для признака `price`.

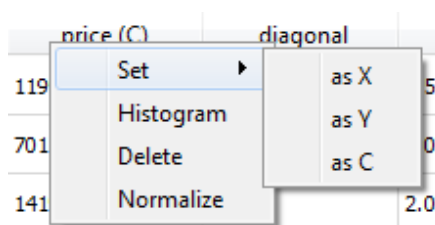


Рисунок 2 – Контекстное меню признака `price`

Контекстное меню содержит следующие пункты:

1. Set — устанавливает особые метки для признака (см. 5.5.2)
 - 1.1. as X — выставить метку X для признака — сделать его осью абсцисс
 - 1.2. as Y — выставить метку Y для признака — сделать его осью ординат
 - 1.3. as C — выставить метку C для признака — использовать признак для задания цвета и формы. (■ ▲ ▼ ●)
2. Histogram — строит гистограмму по выбранному признаку (см. 5.5.1)
3. Delete — удаляет признак из вкладки, в которой вызвано контекстное меню (см. 5.4.1)
4. Normalize — нормализует выбранный признак, добавляя на панель нормализованных данных (см. 5.3.2)

4.6.3 Диалог нормализации

На рисунке 3 показан диалог нормализации. Это окно требует от пользователя выставить значения (см. 4.4,5.3).

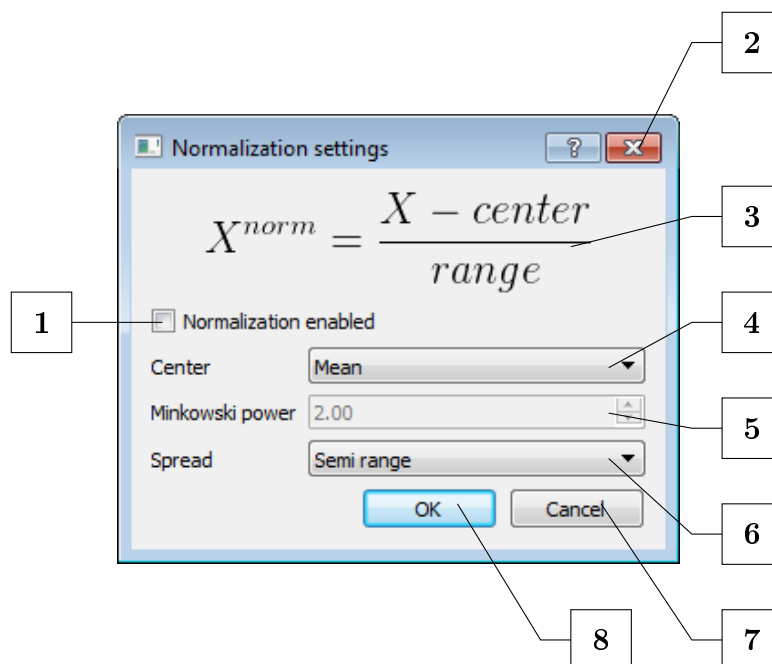


Рисунок 3 – Диалог установки параметров нормализации данных

Цифры на рисунке означают:

1. Переключатель вкл./выкл. нормализацию
2. Кнопка закрытия окна
3. Расчётная формула
4. Выпадающий список для выбора центра нормализации
5. Поле ввода степени Минковского (активно, когда выбран центр Минковского)
6. Выпадающий список для выбора диапазона нормализации
7. Кнопка отмены
8. Кнопка подтверждения ввода

4.6.4 Окно графического вывода

Окно графической информации служит для просмотра различного вида графиков и диаграмм. Такое окно может встретиться пользователю, например при построении гистограммы (раздел 5.5.1), SVD диаграммы (5.5.3) или поля рассеяния (5.5.2).

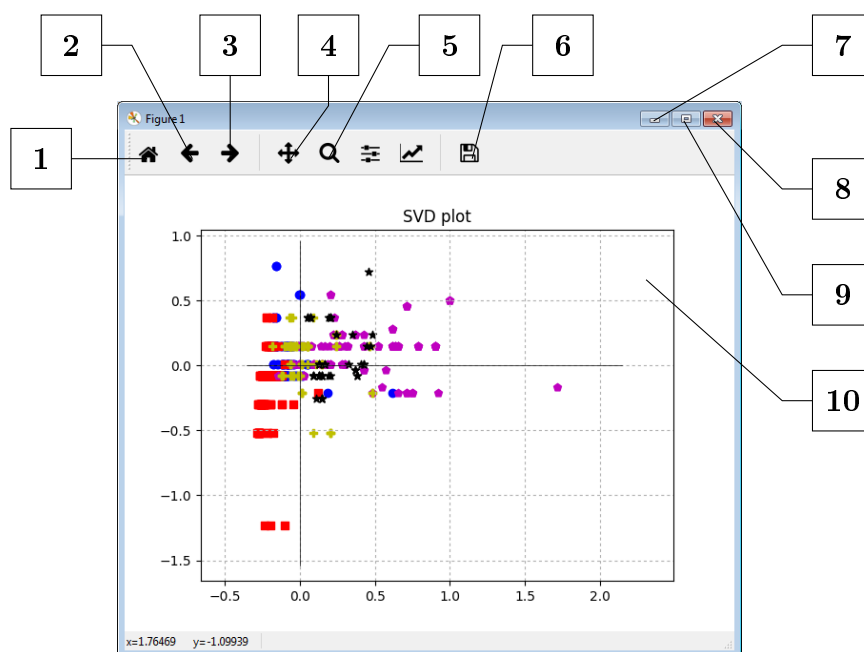


Рисунок 4 – Окно графического вывода

Цифры на рисунке означают:

1. Кнопка восстановления исходного автоматического положения и масштаба
2. Кнопка возврата к предыдущему виду (после масштабирования или смещения)
3. Кнопка возврата к следующему виду (после масштабирования или смещения)
4. Кнопка смещения диаграммы
5. Кнопка масштабирования выбранной области
6. Кнопка сохранения текущего графика в файл
7. Свернуть окно
8. Развернуть окно
9. Закрыть окно
10. Изображение

4.6.5 Окно генерации данных

Окно генерации применяется при работе с синтетическими данными (см. раздел 5.6). Это окно необходимо для ввода информации о значениях характеристик генерируемых данных.

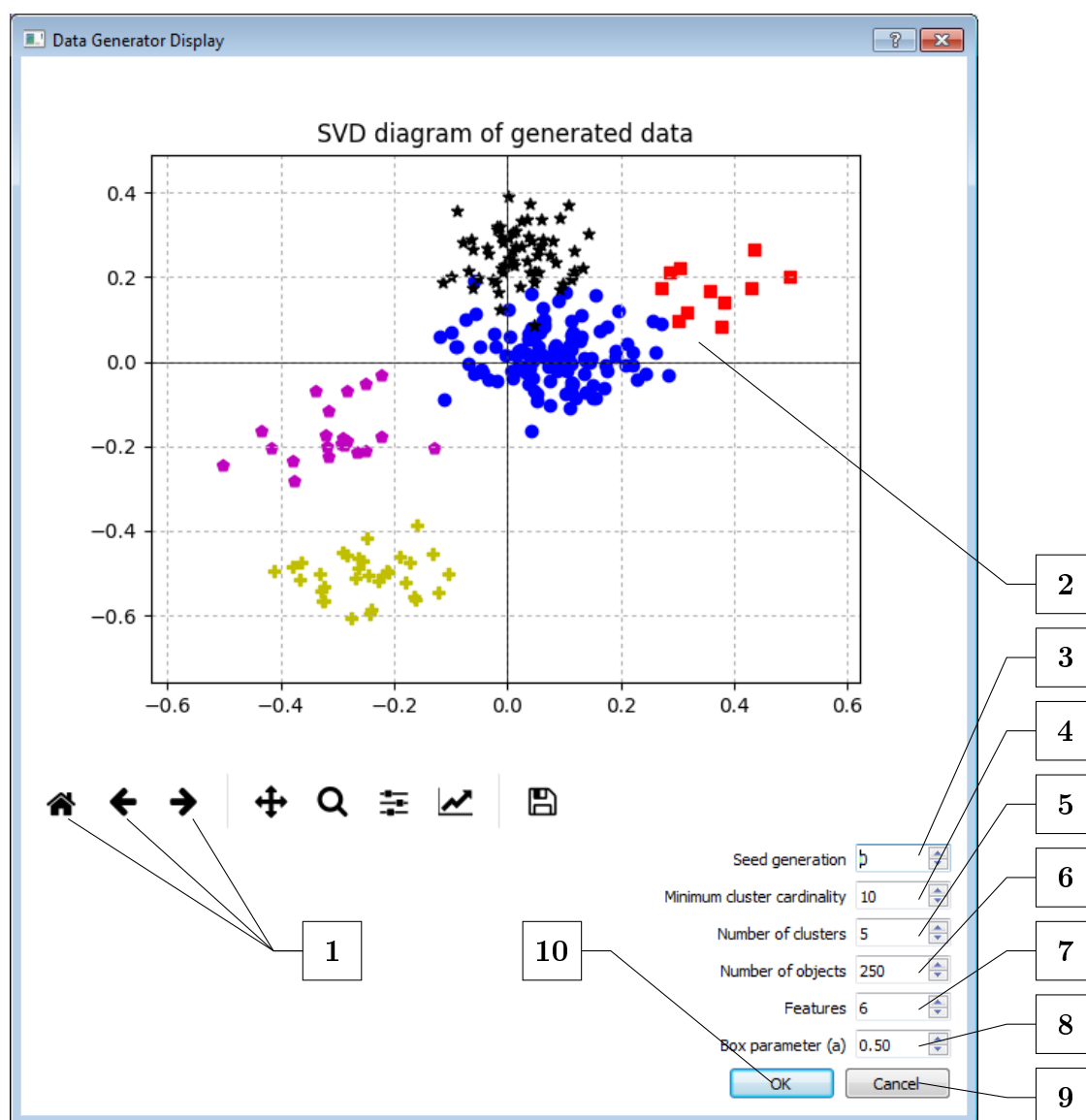


Рисунок 5 – Окно генерации данных

Цифры на рисунке означают:

1. Кнопки управления графиком (см. 4.6.4)

2. Интерактивная графическая иллюстрация результата
3. Поле ввода порождающего значения генератора (seed)
4. Поле ввода минимального числа объектов в кластере
5. Поле ввода числа кластеров
6. Поле ввода числа объектов
7. Поле ввода числа признаков
8. Поле ввода параметра смещения кластеров
9. Кнопка отмены ввода
10. Кнопка подтверждения ввода

4.6.6 Окно запуска кластеризации

В меню каждый алгоритм, реализованный в программе, выделен отдельно. Для задания параметров алгоритма предусмотрены диалоговые окна. В данном разделе будут рассмотрены окна для настройки всех видов алгоритмов, реализованных в программе.

A-Ward

Окно для настройки параметров алгоритма A-Ward изображено на рисунке 6.

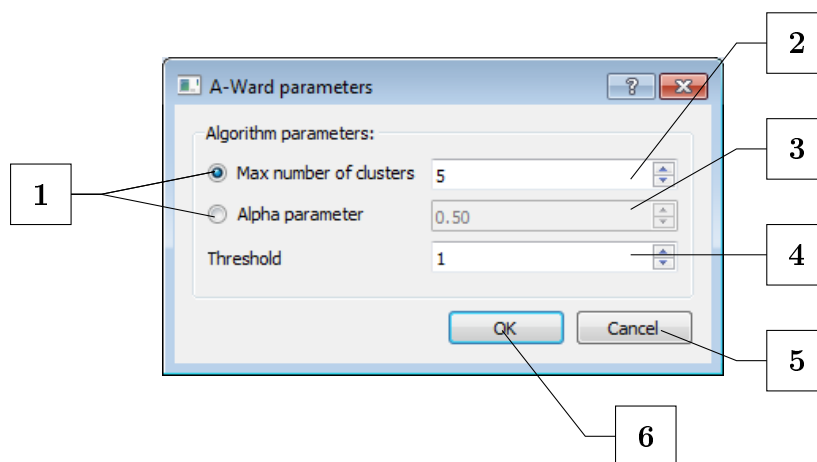


Рисунок 6 – Окно запуска A-Ward

Цифры на рисунке означают:

1. Переключатель критерия остановки алгоритма

2. Поле ввода максимального числа
3. Поле ввода параметра специального критерия
4. Поле ввода порогового значения для отсеечения небольших аномальных кластеров
5. Кнопка отмены ввода
6. Кнопка подтверждения ввода

A-Ward_{p β}

Для запуска алгоритма A-Ward_{p β} используется окно, показанное на рисунке 7.

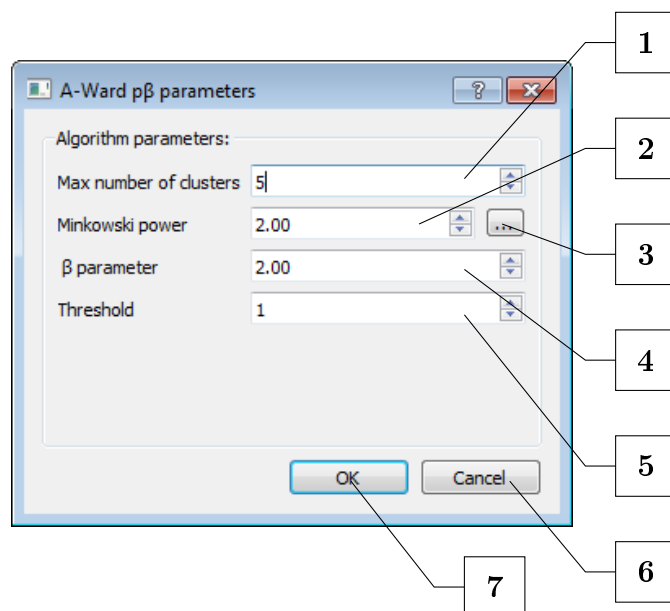


Рисунок 7 – Окно запуска A-Ward_{p β}

Цифры на рисунке означают:

1. Поле ввода максимального числа кластеров
2. Поле ввода степени Минковского
3. Кнопка для запуска автоматического подбора степени Минковского
4. Поле ввода степени весов признаков β
5. Поле ввода порогового значения для отсеечения небольших аномальных кластеров
6. Кнопка отмены ввода

7. Кнопка подтверждения ввода

Bi K-Means R

Для запуска алгоритма Bi K-Means R требуется ввод всего одного параметра, соответствующее окно показано на рисунке 8.

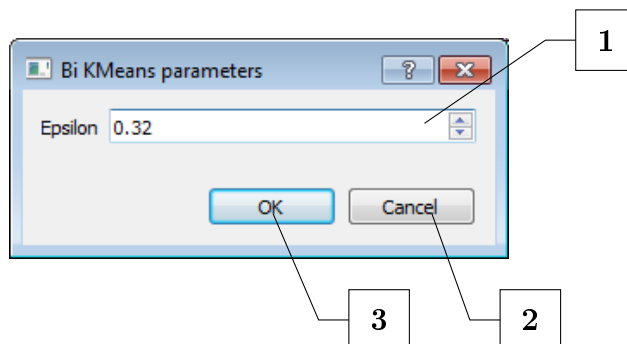


Рисунок 8 – Окно запуска Bi K-Means R

Цифры на рисунке означают:

1. Поле ввода параметра ϵ
2. Кнопка отмены ввода
3. Кнопка подтверждения ввода

dePDDP

Для запуска алгоритма dePDDP не требуется вводить никаких параметров, он запускается сразу же после выбора соответствующего пункта меню.

IK-Means

Окно запуска алгоритма IK-Means такое же как и для A-Ward.

4.6.7 Окно таблицы результатов

Результат кластеризации может быть представлен в табличном виде или в виде текстового отчёта. Эти окна могут быть открыты только после выполнения одного из алгоритмов кластеризации. Окно табличного представления информации о полученных кластерах изображено на рисунке 9. Значение в таблице соответствует среднему значению данного признака в данном кластере. Если это значение существенно больше общего среднего по признаку, то ячейка выделяется красным цветом, если существенно меньше — синим..

	SL	SW	PL	PW	Entities
0	5.006	3.418	1.464	0.244	50
1	6.857	3.091	5.786	2.131	35
2	6.271	2.832	4.842	1.629	31
3	5.380	2.360	3.600	1.060	10
4	5.750	2.817	4.250	1.329	24
Mean	5.843	3.054	3.759	1.199	150

Рисунок 9 – Окно табличного отчёта

Цифры на рисунке означают:

1. Названия признаков
2. Номера кластеров
3. Строка средних значений по всем данным
4. Столбец числа объектов в кластере

Рисунок 10 изображает окно текстового отчёта, которое содержит единственный элемент для просмотра текста.

Intelligent clustering resulted in 5 clusters

Algorithm used: A-Ward p beta with $K^* = 5$; threshold = 1; $p = 2.0$; $\beta = 2.0$; (0.152 s);

Normalization:

- Enabled: True
- Center: Mean
- Spread: Semi range

Clusters characteristics:

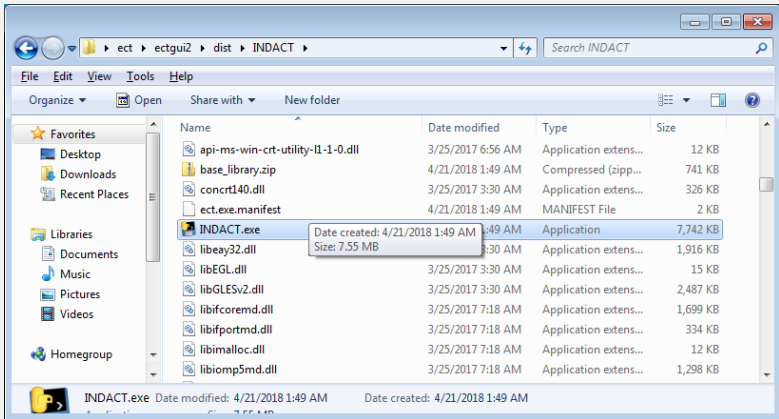
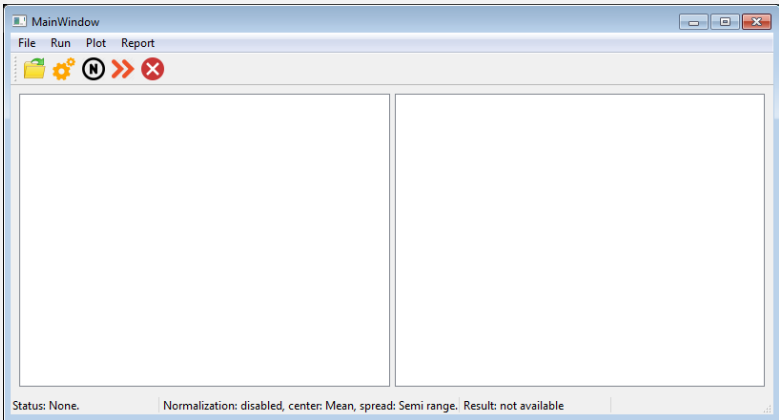
I. Normalized data	SL	SW	PL	PW
Total (150 entities):				
Grand mean:	-0.000	-0.000	0.000	-0.000
Contribution, %:	87.040			
Cluster #1 (50 entities):				

Рисунок 10 – Окно текстового отчёта

5 Описание операций

5.1 Запуск программы

Для работы с программой требуется запустить процесс ОС, который отображает графический интерфейс и взаимодействует с пользователем. Действия этой операции приведены в таблице ниже.

Действие/Описание	Интерфейс
<p><i>1. Запустить бинарный файл программы</i></p> <p>Дважды нажать левой кнопкой мыши (ЛКМ) на значке <code>INDACT.exe</code></p>	
<p><i>2. Дождаться запуска</i></p> <p>Подождать, пока произойдёт инициализация среды выполнения Python. Открытие чёрного консольного окна, означает что установлена отладочная версия программы. Его не следует закрывать. Запуск программы занимает не более 3-4 сек.</p>	


5.2 Загрузка исходных данных

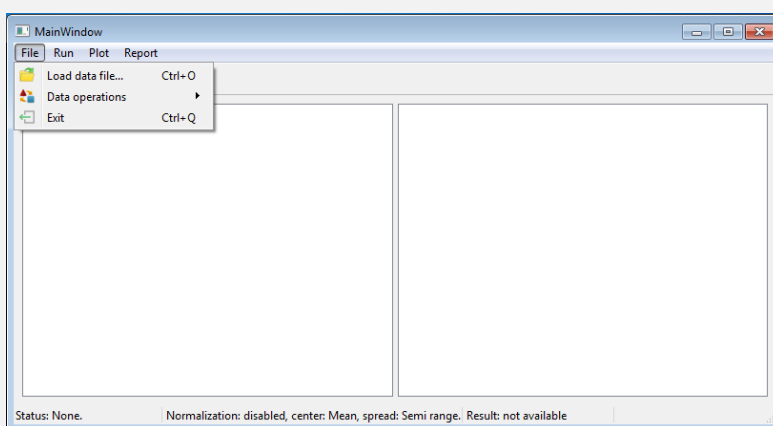
Загрузка данных необходима для того, чтобы подать программе файл, который содержит таблицу данных. Формат файла должен удовлетворять набору требований, перечисленных в разделе [4.2 Требования к файлу исходных данных](#).

Действие/Описание

Интерфейс

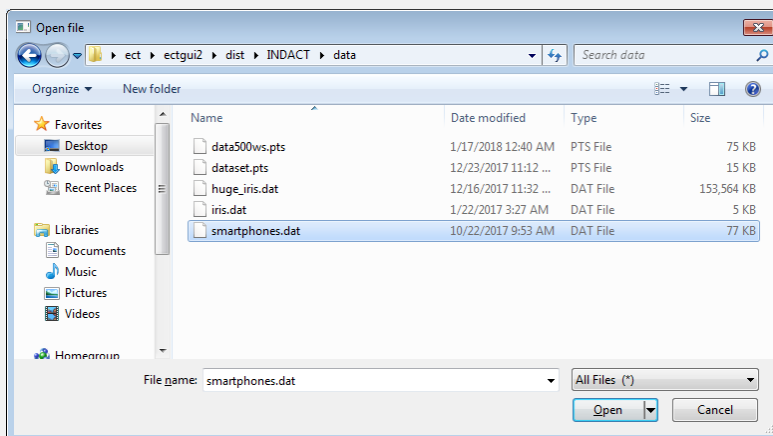
1. Открыть диалог загрузки файла

Для открытия диалога загрузки файла необходимо последовательно нажать в главном меню пункты **File** ⇒ **Load data file**. Того же результата можно достичь нажатием иконки  (см. рисунок 1 обозначение 10)



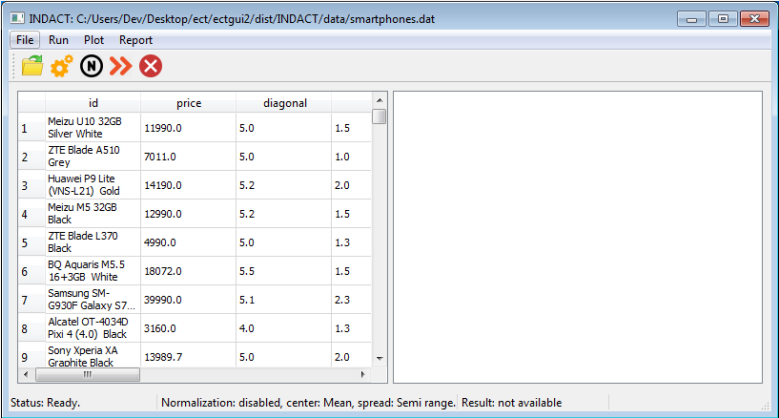
2. Выбрать текстовый файл с данными

В стандартном файловом диалоге необходимо выбрать загружаемый файл и нажать кнопку **Open**. Например, для загрузки демонстрационного примера следует выбрать файл `INDACT/data/smartphones.dat`



3. Проверить загрузку исходного файла

После выполнения предыдущего пункта будет произведена загрузка файла и отображение его содержимого в виде таблицы в интерфейсе программы. Пользователю следует убедиться, что загружен правильный файл, объекты и признаки отображаются верно. На рисунке справа показан загруженный файл `smartphones.dat`




5.3 Нормализация данных

5.3.1 Установка параметров нормализации

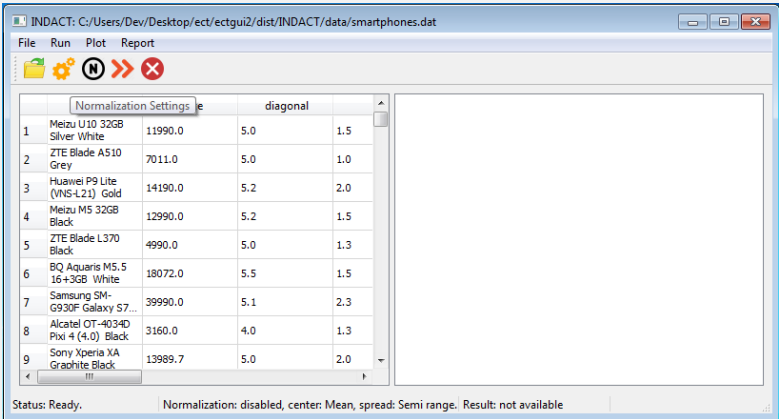
Назначение и параметры нормализации описаны в разделе 4.4 Нормализация данных, примеры установки различных настроек нормализации описаны в разделе 7.1.1.

Действие/Описание

1. Открыть диалог нормализации

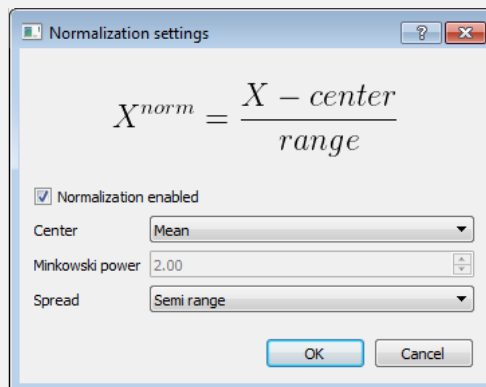
Диалог нормализации можно открыть щелчком ЛКМ на иконке  (см. рисунок 1 обозначение 11)

Интерфейс



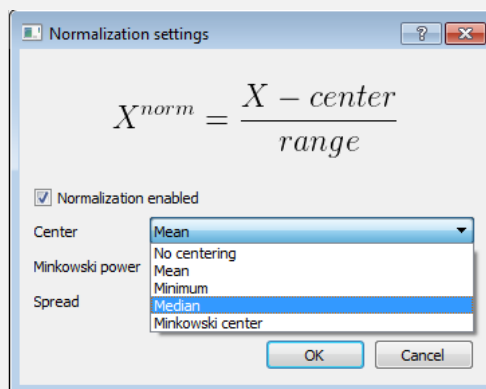
2. Включить нормализацию

Чтобы включить нормализацию, требуется установить флажок “Normalization enabled” в открывшемся окне настроек нормализации.



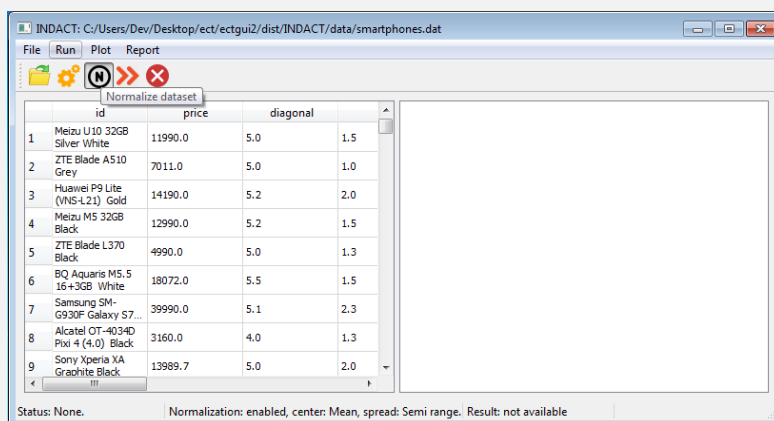
3. Выставить параметры

Для нормализации данных необходимо задать центр и диапазон нормализации. Эти параметры выбираются из выпадающих списков. Нажать **Ok**.



4. Убедиться что нормализация включена

После применения настроек нормализации следует убедиться, что кнопка нормализации успешно включилась (см. рисунок 1 обозначение 12).



5.3.2 Нормализация одного признака

После настройки параметров нормализации необходимо выбрать, какие признаки требуется нормализовать. Только выбранные признаки будут участвовать в кластеризации. В программе предусмотрено четыре возможности для выбора признаков: выбор по одному, выбор нескольких сразу, удаление по одному, удаление сразу нескольких.

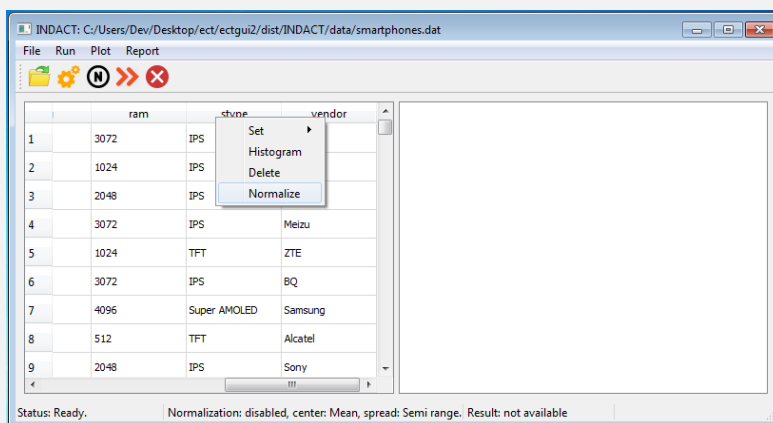
Действие/Описание

Интерфейс

1. Выбрать

признак для нормализации

Для выбора одного признака необходимо найти столбец признака на панели исходных данных (рисунок 1 обозначение 20) и нажать на нем правой кнопкой мыши (ПКМ). В контекстном меню выбрать пункт **Normalize**. На примере показана операция нормализации признака `stype`.



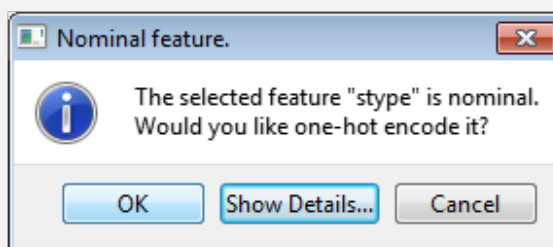
2. (При необходимости)

Подтвердить

нормализацию

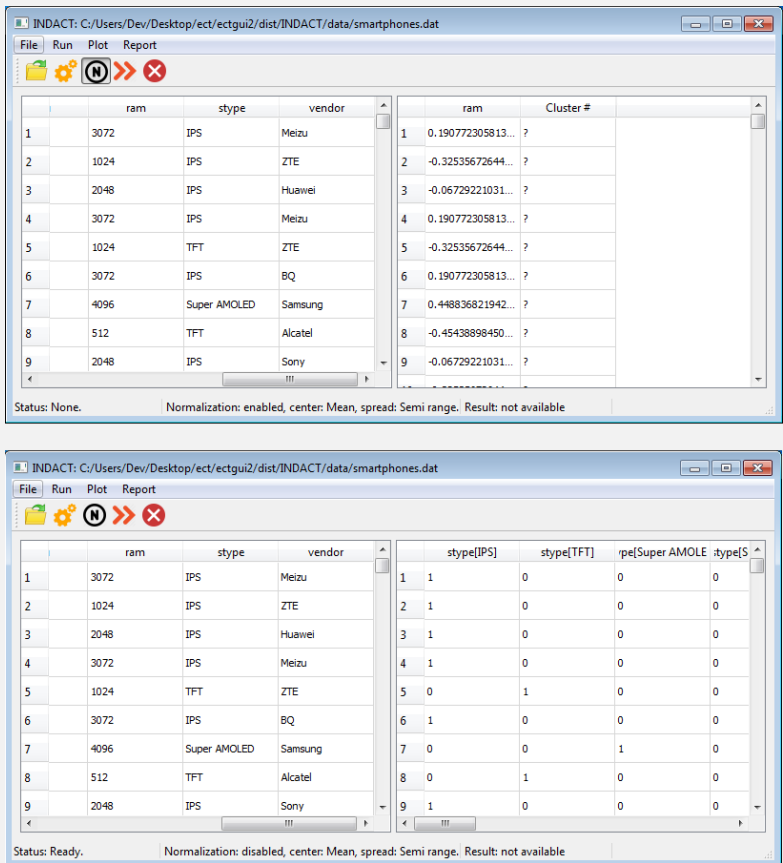
категориального признака

Если был выбран категориальный признак (в примере `stype`), то программа запросит подтверждение разложения признака на бинарные. В случае согласия произойдет добавление бинарных признаков, отвечающих за наличие каждого из значений категориального признака и их нормализация.



3. Просмотр
вида таблицы

После выбора признака, он будет перенесён из вкладки на панель нормализованных данных и к нему будут применены выбранные настройки нормализации. Дополнительно на панели нормализованных данных будет отображён столбец “Cluster#”, который будет оставаться заполненным символами “?” до тех пор, пока не будет выполнен шаг кластеризации (см. верхний рисунок, нормализация признака `rom`). Сказанное выше справедливо и для номинального признака (например `stype`), но стоит иметь в виду что соответствующие бинарные признаки будут названы `stype[значение признака]` (как показано на нижнем рисунке)




5.3.3 Нормализация нескольких признаков сразу

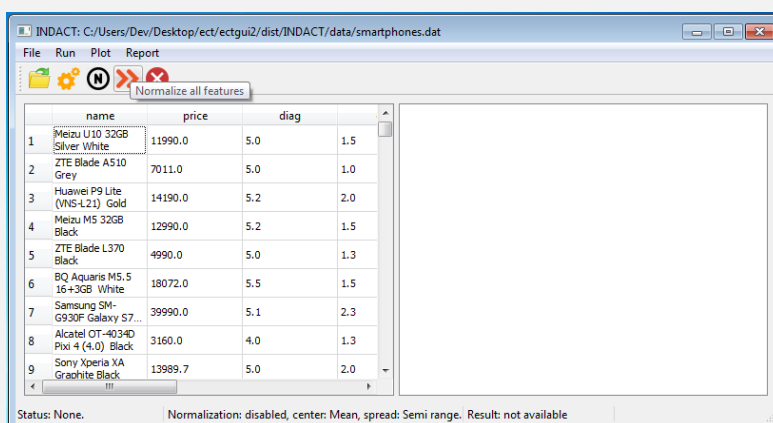
Если признаков много и нормализовать их по одному долго, то можно воспользоваться функцией нормализации нескольких признаков сразу.

Действие/Описание

Интерфейс

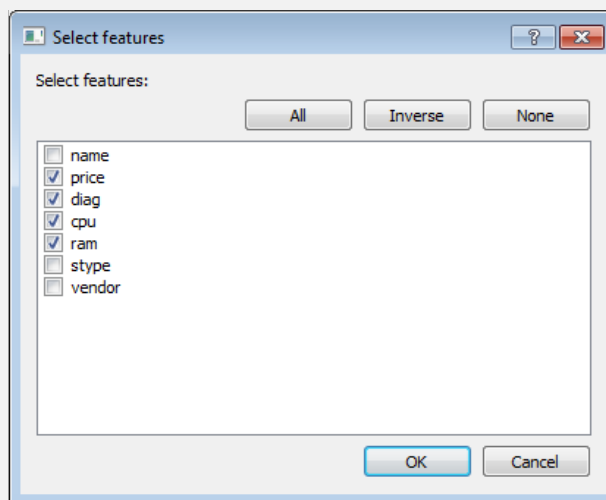
1. Запустить нормализацию нескольких признаков

Для запуска нормализации нескольких признаков сразу, требуется нажать иконку . Откроется окно выбора признаков.



2. Выбрать признаки для нормализации

В открытом окне выбрать те признаки, которые требуется нормализовать. По умолчанию отмечены признаки, которые не являются категориальными. Если требуется нормализовать в том числе категориальные, их следует отметить. Подтвердить выбор признаков, нажав кнопку **Ok**. Если имеется хотя бы один категориальный признак, то программа запросит подтверждение разложения признака по количеству уникальных значений. В случае согласия программа представит номинальный признак с помощью бинарных.



3. Посмотреть результат

После нормализации признаков результат будет отображен на панели нормализованных данных (см. рисунок 1, обозначение 21)

	name	price	diag	
1	Meizu U10 32GB Silver White	11990.0	5.0	1.5
2	ZTE Blade A510 Grey	7011.0	5.0	1.0
3	Huawei P9 Lite (VNS-L21) Gold	14190.0	5.2	2.0
4	Meizu M5 32GB Black	12990.0	5.2	1.5
5	ZTE Blade L370 Black	4990.0	5.0	1.3
6	BQ Aquaris M5.5 16+3GB White	18072.0	5.5	1.5
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 Grachite Black	13989.7	5.0	2.0

	price	diag	cpu	
1	11990.0	5.0	1.5	3072
2	7011.0	5.0	1.0	1024
3	14190.0	5.2	2.0	2048
4	12990.0	5.2	1.5	3072
5	4990.0	5.0	1.3	1024
6	18072.0	5.5	1.5	3072
7	39990.0	5.1	2.3	4096
8	3160.0	4.0	1.3	512
9	13989.7	5.0	2.0	2048

Status: Ready. Normalization: disabled, center: Mean, spread: Semi range. Result: not available

5.4 Отбор признаков

5.4.1 Удаление одного признака

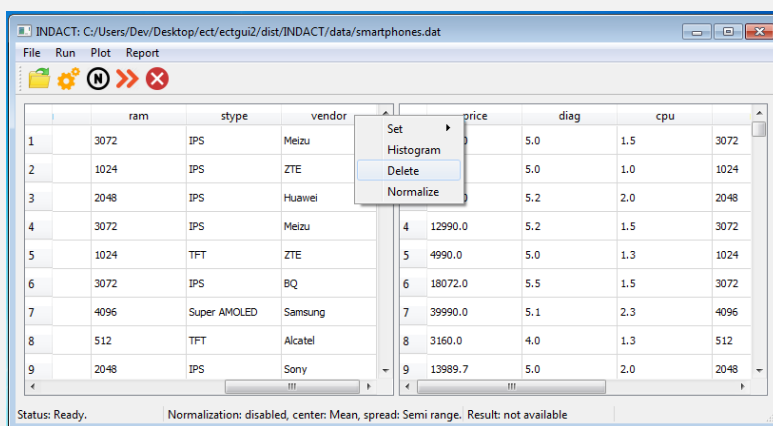
Как было отмечено выше, программа позволяет удалять отдельные признаки как с панели нормализованных данных, так и с панели исходных данных. Эта функция может быть применена для исключения определённых признаков.

Действие/Описание

Интерфейс

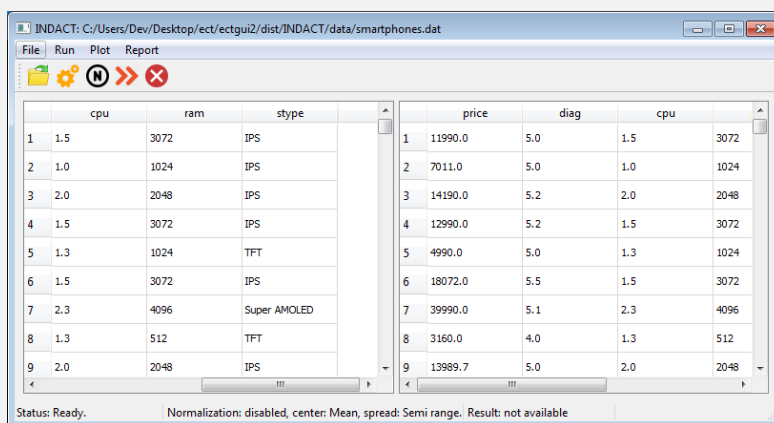
1. Выбрать признак для удаления

Для удаления одного признака необходимо найти столбец признака в нужной вкладке и нажать на нём ПКМ. В контекстном меню выбрать пункт **Delete**. Рассмотрим удаление на примере признака **vendor**. Контекстное меню, открытое после нажатия на заголовке **vendor** показано на рисунке справа.



2. Посмотреть результат

В результате выполнения операции выбранный признак будет удалён только из панели исходных данных, но останется на второй панели.




5.4.2 Удаление нескольких признаков сразу

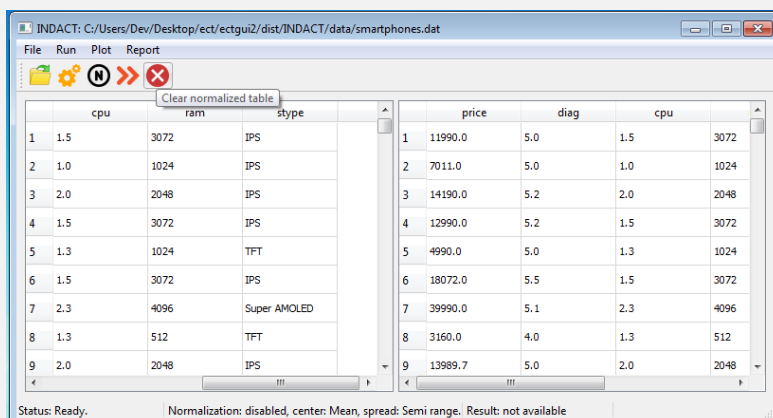
Если требуется полностью очистить панель нормализованных данных или удалить большое количество признаков, то следует воспользоваться функцией, описанной ниже. Функция удаления нескольких признаков сразу может быть использована для быстрого исключения лишних признаков.

Действие/Описание

Интерфейс

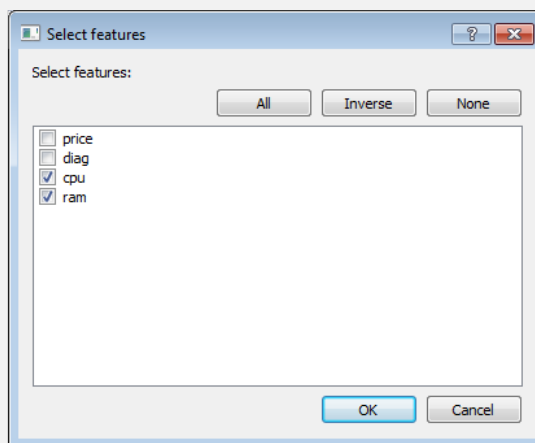
1. Запустить удаление

Функция удаления нескольких признаков вызывается щелчком ЛКМ на иконке  (см. рисунок 1 обозначение 14).



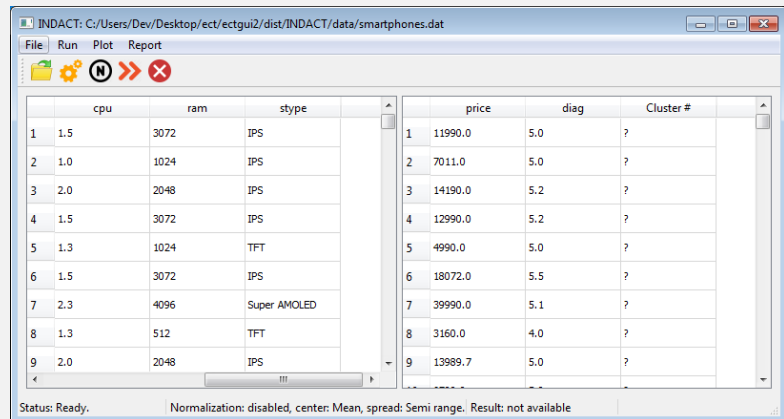
2. Выбрать удаляемые признаки

Выбор удаляемых признаков производится при помощи выставления соответствующих флажков. Все отмеченные признаки будут удалены из панели нормализованных данных. Когда выбор завершён, нажать кнопку **Ok**.



3. Посмотреть результат

В результате выполнения операции выбранные признаки будут удалены с панели нормализованных данных.



	cpu	ram	stype	price	diag	Cluster #
1	1.5	3072	IPS	11990.0	5.0	?
2	1.0	1024	IPS	7011.0	5.0	?
3	2.0	2048	IPS	14190.0	5.2	?
4	1.5	3072	IPS	12990.0	5.2	?
5	1.3	1024	TFT	4990.0	5.0	?
6	1.5	3072	IPS	18072.0	5.5	?
7	2.3	4096	Super AMOLED	39990.0	5.1	?
8	1.3	512	TFT	3160.0	4.0	?
9	2.0	2048	IPS	13989.7	5.0	?

Status: Ready. Normalization: disabled, center: Mean, spread: Semi range. Result: not available

5.5 Визуализация

5.5.1 Построение гистограммы признака

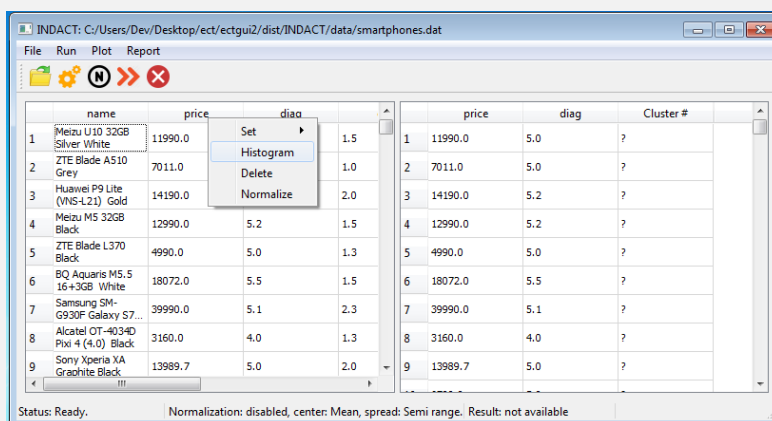
В качестве первичного инструмента анализа программа предлагает возможность построения гистограммы выбранного признака.

Действие/Описание

Интерфейс

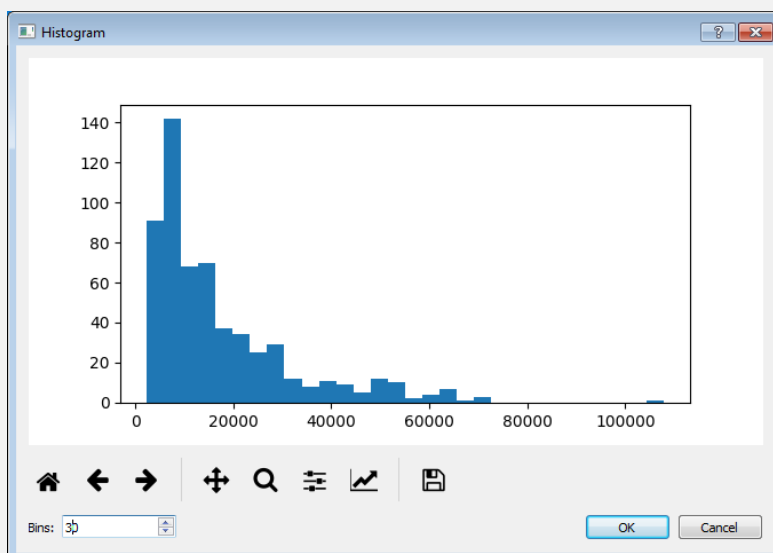
1. Выбрать признак

Для построения гистограммы необходимо определиться с признаком, по которому будет построена гистограмма. Для этого надо щёлкнуть ПКМ на названии столбца соответствующего признака. В контекстном меню выбрать пункт **Histogram**. На примере показано построение гистограммы по признаку **price**



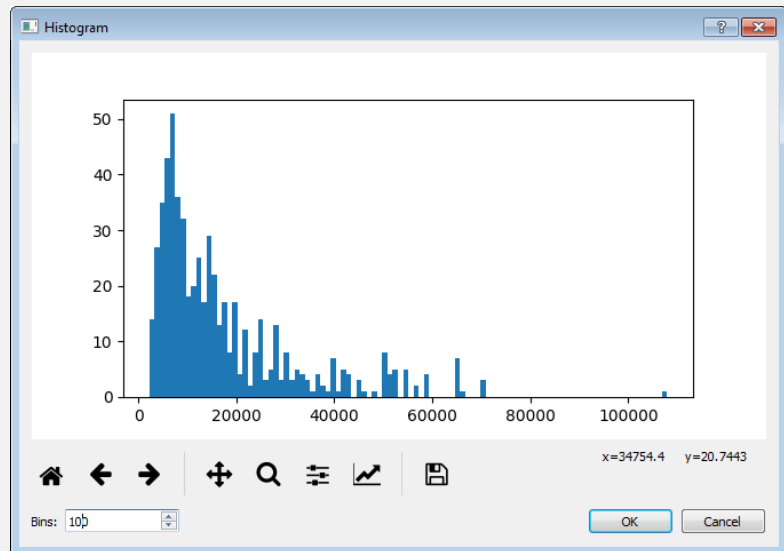
2. Отрегулировать количество бинов

Как известно, вид гистограммы может зависеть от числа интервалов, на которые разбиваются область допустимых значений - бинов. Поэтому в программе предусмотрена регулировка этого параметра. В нижнем поле ввода можно выставить требуемое число интервалов.



3. Посмотреть результат

Результат построения гистограммы будет доступен сразу же в том же окне. На рисунке справа показана гистограмма, построенная для ненормализованного признака `price` при количестве интервалов, равном 100. Окно отображения гистограммы обладает всеми стандартными кнопками для управления визуализациями, в том числе для сохранения в файл.



5.5.2 Построение поля рассеяния (scatter plot)

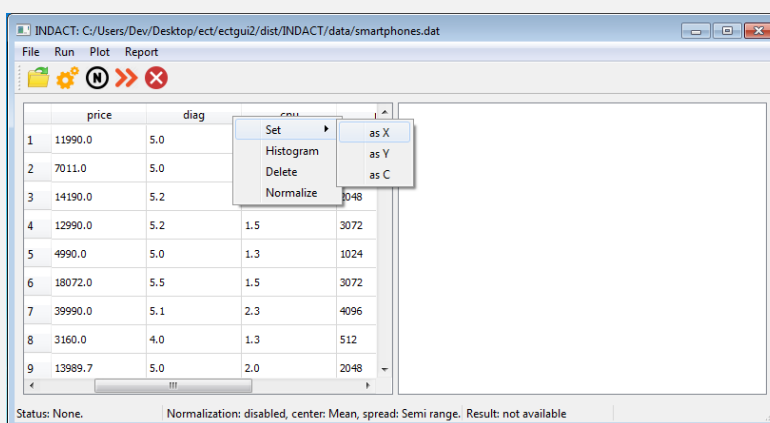
Для первичного анализа структуры данных или результатов кластеризации в программе предусмотрена функция построения поля рассеяния по меткам на выбранных признаках. Метка — вспомогательный символ, присваиваемый пользователем для определённого признака. Предусмотрено 3 вида меток: “X”, “Y”, “C”. Первый вид означает что отмеченный признак будет соответствовать координатам объекта по оси абсцисс, второй — по оси ординат, а третий, что цвет (*Color*) точки будет выбираться в соответствии со значением отмеченного признака

Действие/Описание

Интерфейс

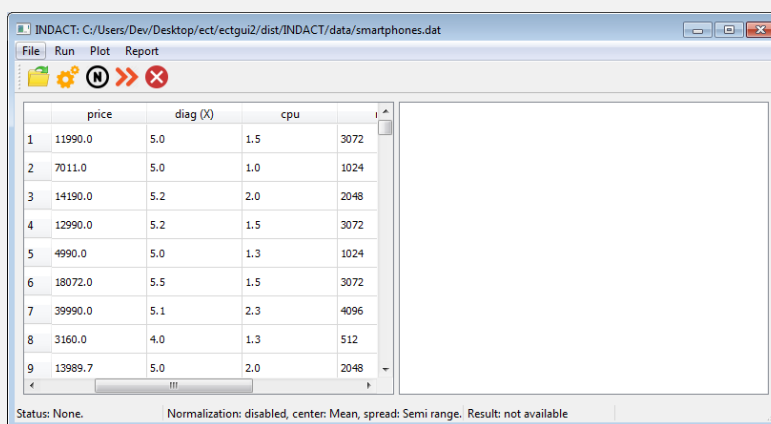
1. Выбрать признак по оси X

Для построения поля рассеяния требуется задать признаки по осям абсцисс и ординат. Чтобы отметить признак, соответствующий оси абсцисс, требуется нажать на его названии ПКМ и в контекстном меню выбрать **Set** ⇒ **as X**



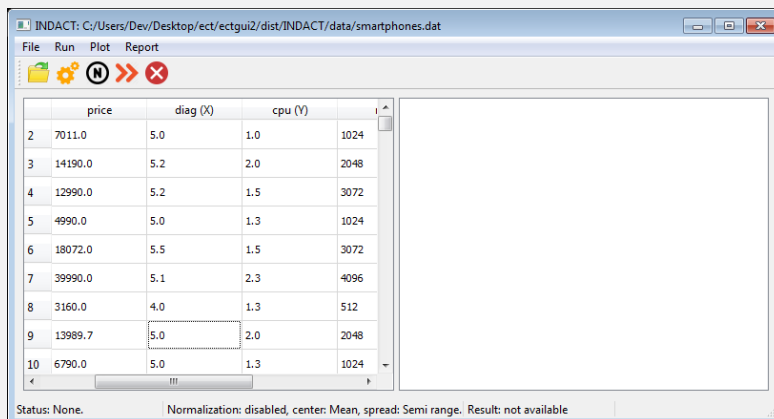
2. Посмотреть результат

После установки маркера “X” к имени соответствующего признака добавиться “(X)”. На примере показано добавление метки “X” к признаку *diag*.



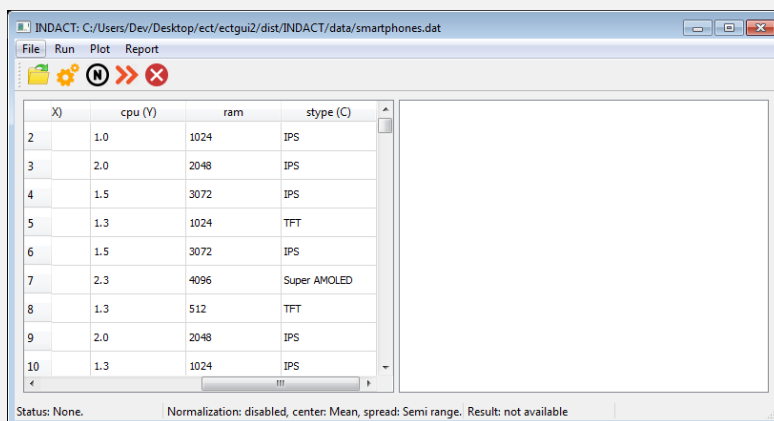
3. Выбрать признак по оси Y

Аналогично пунктам 1,2. В примере указан признак *cpu*.



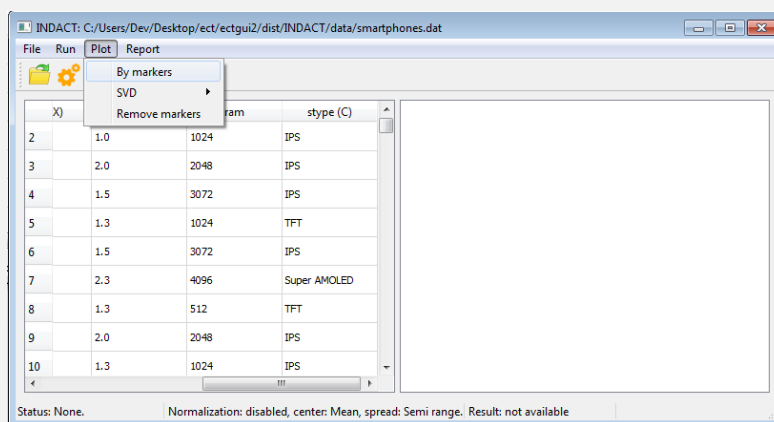
4. Выбрать признак, отвечающий за цвет (опционально)

Для того чтобы задать, какой признак будет определять цвет точек на диаграмме, необходимо выставить маркер "C". Для этого выбрать в контекстном меню выбрать **Set** ⇒ **as C**.



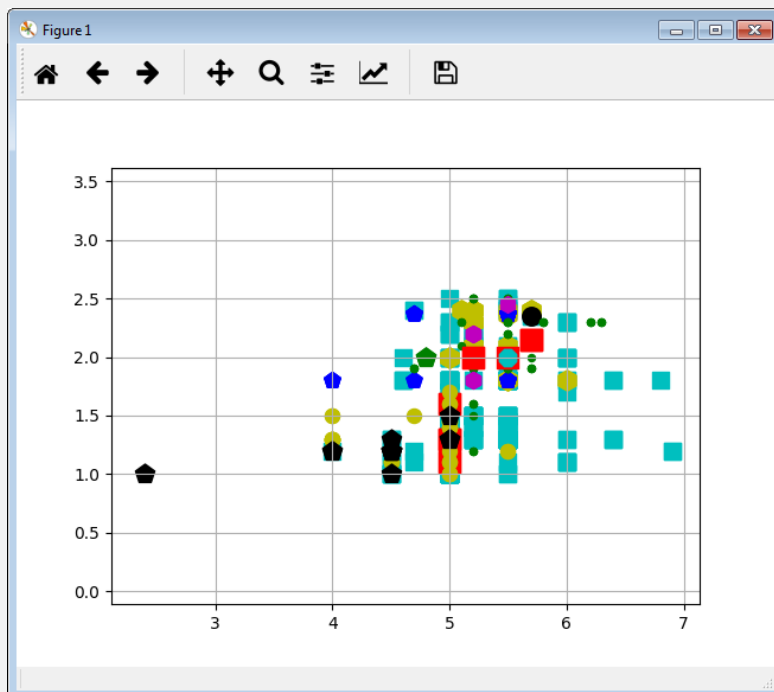
5. Построить scatter plot

В главном меню выбрать **Plot** ⇒ **By markers**



6. Посмотреть результат

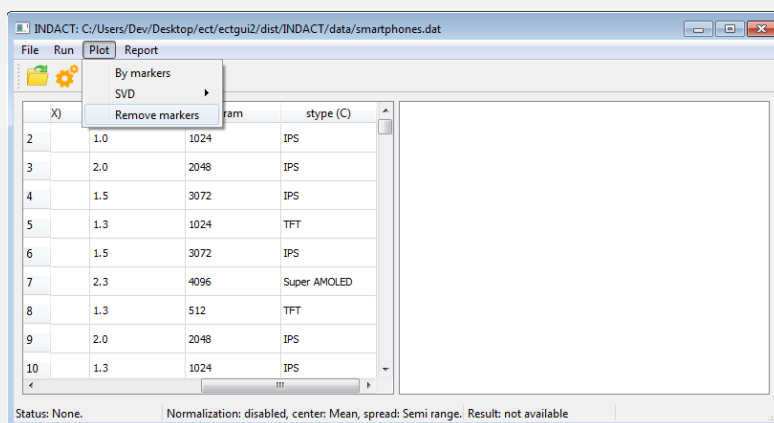
В новом окне откроется построенная красивая и цветастая диаграмма. Как и прочие окна отображения графической информации, окно отображения scatter plot имеет кнопки для управления видом и сохранения изображения в файл (подробнее см. рисунок 4, обозначения 1-6). На примере показана диаграмма, у которой по оси X отложен признак `diag`, по оси Y — `cpu`, цвет определяется признаком `stype`.



7. Удалить

метки (опционально)

В главном меню выбрать **Plot** ⇒ **Delete All Markers**. При этом отметки “X”, “Y” и “C” будут удалены.



5.5.3 Построение SVD диаграммы

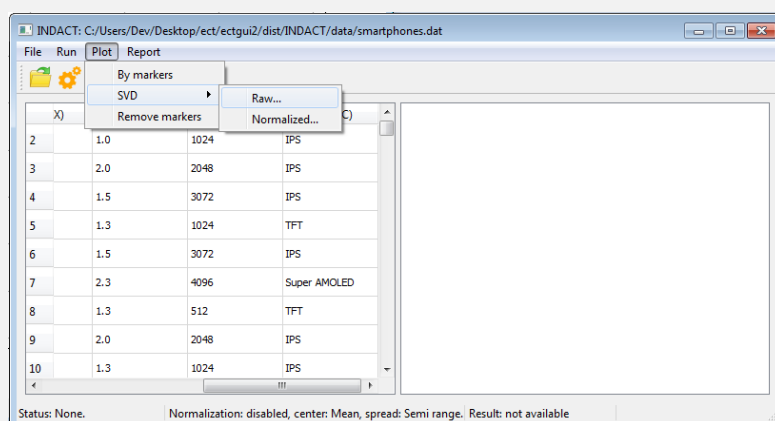
Для интегрального представления структуры данных предусмотрена функция построения SVD диаграммы. Имеется возможность построения SVD диаграммы по нормализованным и не нормализованным данным.

Действие/Описание

Интерфейс

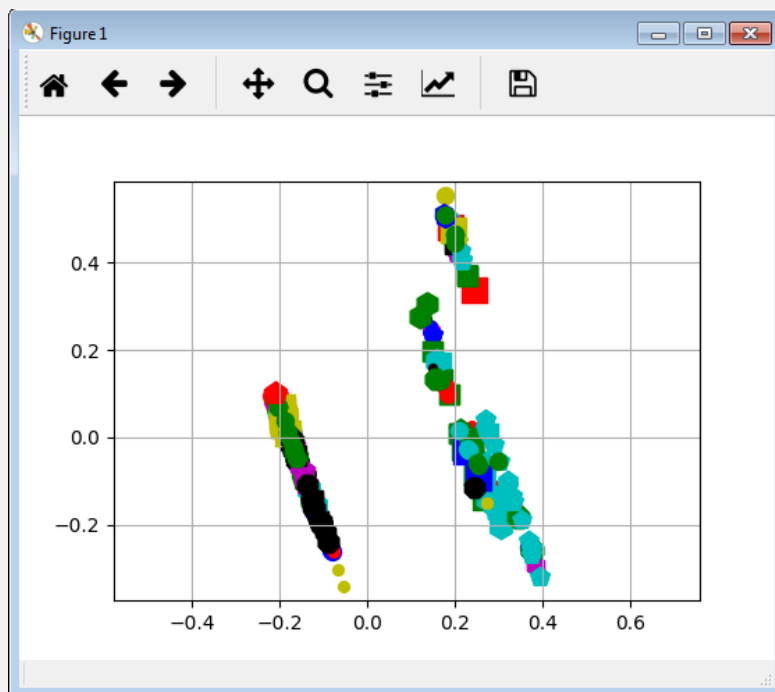
1. Построить SVD диаграмму

Для построения SVD диаграммы в главном меню выбрать **Plot** ⇒ **SVD** ⇒ **Normalized** или **Raw** для построения диаграммы по нормализованным и не нормализованным данным соответственно.



2. Посмотреть результат

Построенная диаграмма откроется в новом окне, как показано на рисунке справа.



5.6 Генерация синтетических данных

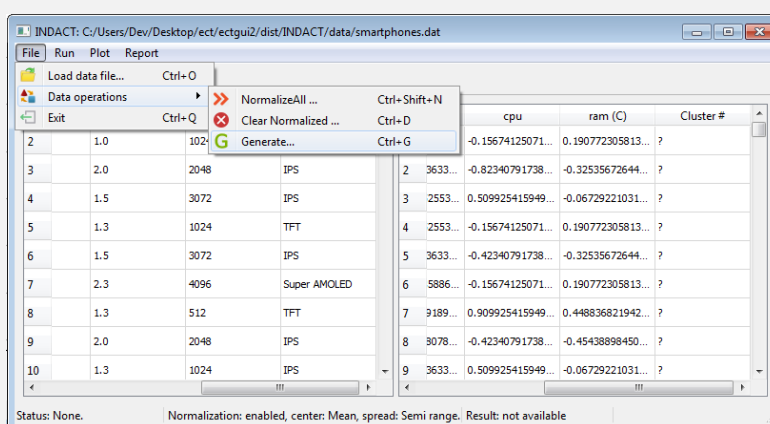
Для генерирования искусственных данных необходимо вызвать диалог настройки параметров, указать все необходимые величины и сохранить сгенерированные данные в файл.

Действие/Описание

Интерфейс

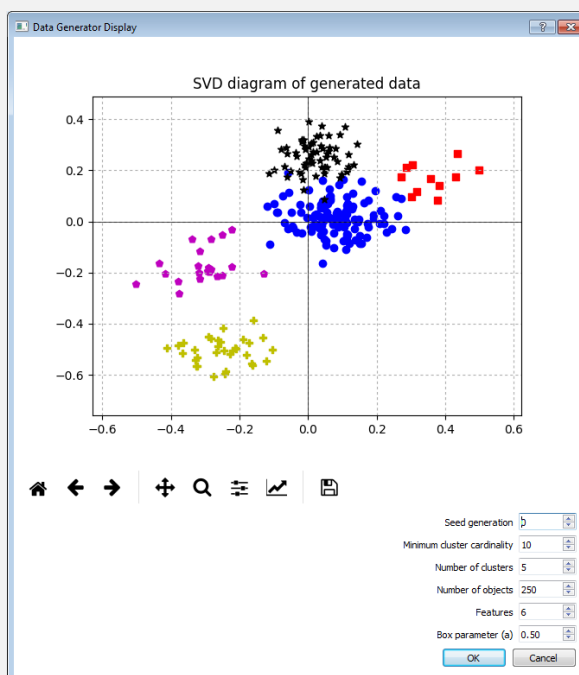
1. Открыть диалог генерации данных

Чтобы открыть диалог загрузки данных, необходимо в главном меню программы выбрать **File** ⇒ **Data operations** ⇒ **Generate**.



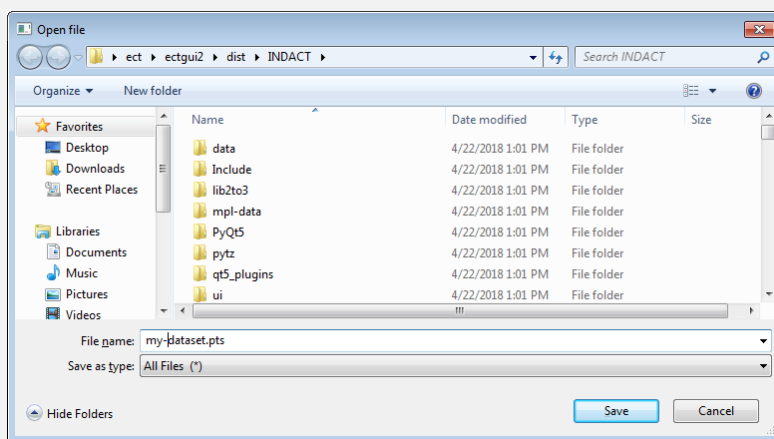
2. Указать параметры данных

В открывшемся диалоге необходимо указать параметры данных, по которым будет производиться генерация. Подробнее о параметрах генерации см. [3]. В верхней части диалогового окна отображается динамическая информирующая диаграмма. Когда все параметры будут введены, нажать кнопку **OK**.



3. Указать имя файла

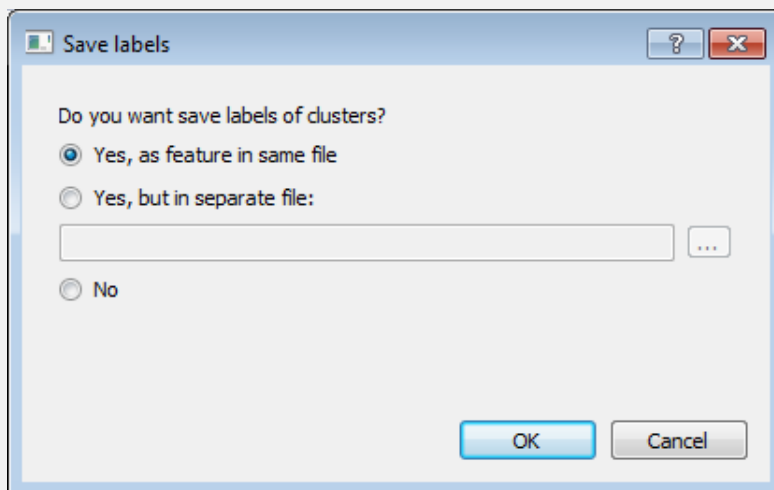
В файловом диалоге указать имя файла, куда будут сохранены сгенерированные данные. Нажать кнопку **Ok**



4. Выбрать

как сохранить кластеры

Принадлежность кластера определяется по целочисленными меткам, присвоенным каждому объекту. Для задания способа сохранения меток откроется окно, как показано на рисунке справа. На выбор доступны опции сохранения меток как отдельный признак в том же файле, в отдельном файле или их можно вообще не сохранять. Выбрав нужную опцию, нажать **Ok**. Файлы будут сохранены соответственно выбранным настройкам.



5.7 Запуск кластеризации

Для определения принадлежности объектов кластерам, требуется установить параметры кластеризации и запустить алгоритм.

Действие/Описание

Интерфейс

1. Открыть

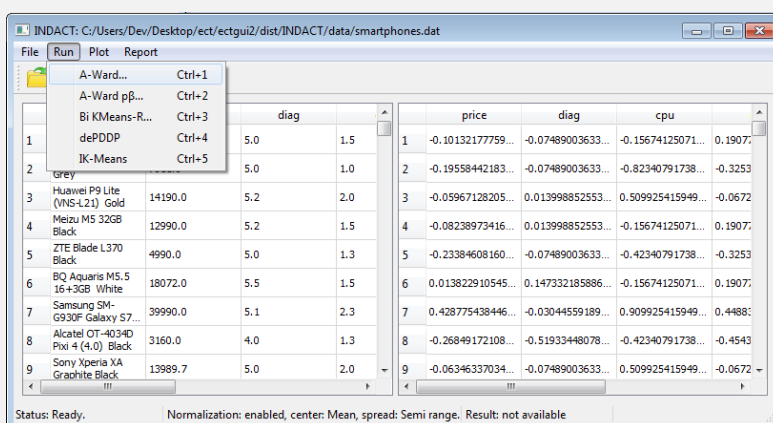
диалог выбора параметров

Для запуска алгоритма кластеризации первым делом требуется открыть диалог настройки параметров. Для этого выбрать в главном меню

Run

⇒ **<Алгоритм>**, где

<Алгоритм> соответствует названию одного из реализованных алгоритмов, например, A-Ward. Список всех реализованных алгоритмов можно увидеть на рисунке.

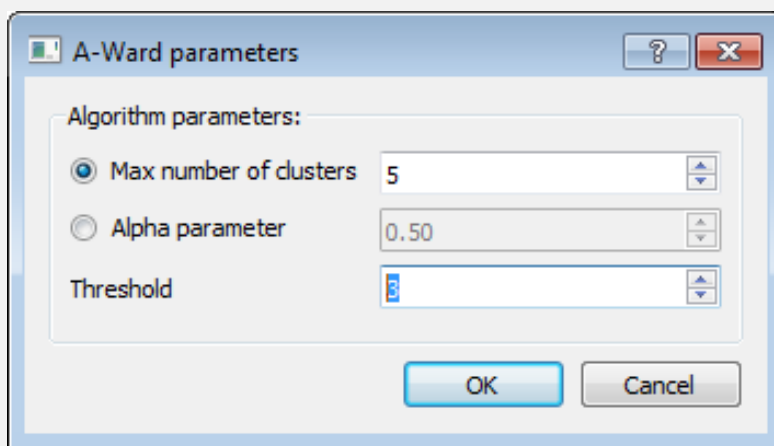


2. Настроить

параметры алгоритма

Для большинства алгоритмов требуется указать некоторые управляющие параметры, для справки по соответствующим диалоговым окнам см. раздел. 4.6.6 Окно запуска кластеризации. После задания параметров нажать кнопку

Ok.



3. Дождаться результатов кластеризации

Сразу после нажатия кнопки **Ok** будет запущена работа алгоритма. Когда кластеризация будет завершена, в строке состояния будет выведена краткая информация о алгоритме и времени работы. В то же время в столбце **Cluster #** будут проставлены метки кластеров. Каждая метка соответствует номеру кластера, которому принадлежит объект.

	name	price	diag		cpu	ram	Cluster #
1	Meizu U10 32GB Silver White	11990.0	5.0	1.5	2 3633...	-0.82340791738...	-0.32535672644... 3
2	ZTE Blade A510 Grey	7011.0	5.0	1.0	3 2553...	0.509925415949...	-0.06729221031... 4
3	Huawei P9 Lite (VNS-L21) Gold	14190.0	5.2	2.0	4 2553...	-0.15674125071...	0.190772305813... 1
4	Meizu M5 32GB Black	12990.0	5.2	1.5	5 3633...	-0.42340791738...	-0.32535672644... 3
5	ZTE Blade L370 Black	4990.0	5.0	1.3	6 5886...	-0.15674125071...	0.190772305813... 1
6	BQ Aquaris M5.5 16+3GB White	18072.0	5.5	1.5	7 9189...	0.909925415949...	0.448836821942... 0
7	Samsung SM-G930F Galaxy S7...	39990.0	5.1	2.3	8 3078...	-0.42340791738...	-0.45438898450... 3
8	Alcatel OT-4034D Pixi 4 (4.0) Black	3160.0	4.0	1.3	9 3633...	0.509925415949...	-0.06729221031... 4
9	Sony Xperia XA Gracchite Black	13989.7	5.0	2.0	10 3633...	-0.42340791738...	-0.32535672644... 3

Status: None. Normalization: enabled, center: Mean, spread: Semi range. Result: (1.03 s) A-Ward with threshold = 3; K* = 5;

5.8 Генерация отчёта

Результаты кластеризации удобно анализировать по сгенерированному отчёту. Программа INDACT предлагает два вида отчёта: текстовый и табличный.

5.8.1 Текстовый отчёт

Действие/Описание

Интерфейс

1. Сгенерировать отчёт

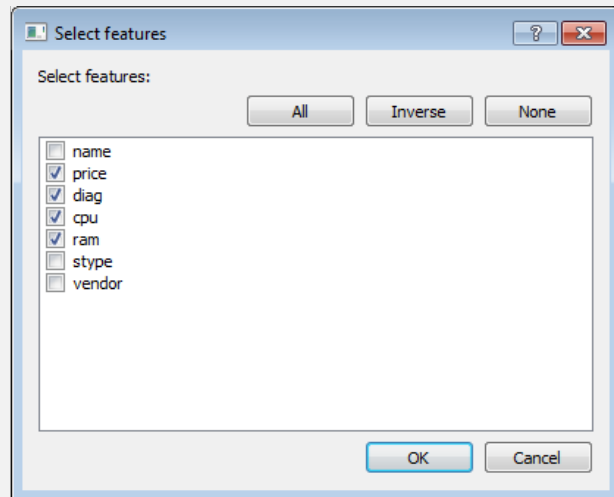
Для генерации текстового отчёта в главном меню выбрать **Report** ⇒ **Text**.

	name	price	diag		cpu	ram	Cluster #
1	Meizu U10 32GB Silver White	11990.0	5.0	1.5	2 3633...	-0.82340791738...	-0.32535672644... 3
2	ZTE Blade A510 Grey	7011.0	5.0	1.0	3 2553...	0.509925415949...	-0.06729221031... 4
3	Huawei P9 Lite (VNS-L21) Gold	14190.0	5.2	2.0	4 2553...	-0.15674125071...	0.190772305813... 1
4	Meizu M5 32GB Black	12990.0	5.2	1.5	5 3633...	-0.42340791738...	-0.32535672644... 3
5	ZTE Blade L370 Black	4990.0	5.0	1.3	6 5886...	-0.15674125071...	0.190772305813... 1
6	BQ Aquaris M5.5 16+3GB White	18072.0	5.5	1.5	7 9189...	0.909925415949...	0.448836821942... 0
7	Samsung SM-G930F Galaxy S7...	39990.0	5.1	2.3	8 3078...	-0.42340791738...	-0.45438898450... 3
8	Alcatel OT-4034D Pixi 4 (4.0) Black	3160.0	4.0	1.3	9 3633...	0.509925415949...	-0.06729221031... 4
9	Sony Xperia XA Gracchite Black	13989.7	5.0	2.0	10 3633...	-0.42340791738...	-0.32535672644... 3

Status: None. Normalization: enabled, center: Mean, spread: Semi range. Result: (1.03 s) A-Ward with threshold = 3; K* = 5;

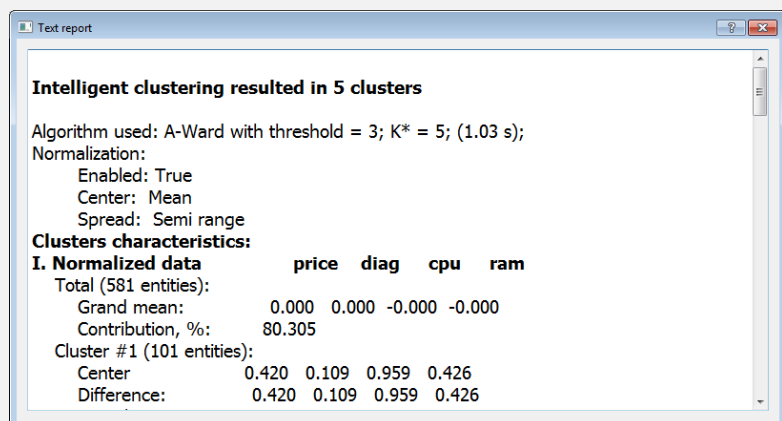
2. Выбрать признаки

Выбрать признаки, которые будут включены в текстовый отчёт. Для этого отметить флажки соответствующие тем признакам, которые требуется указать в отчёте.

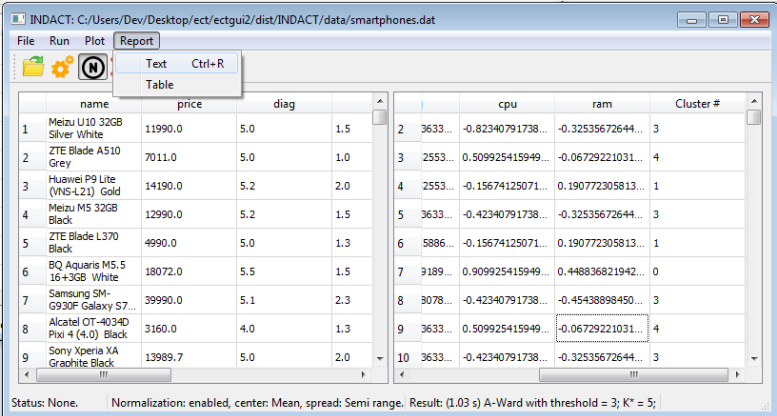
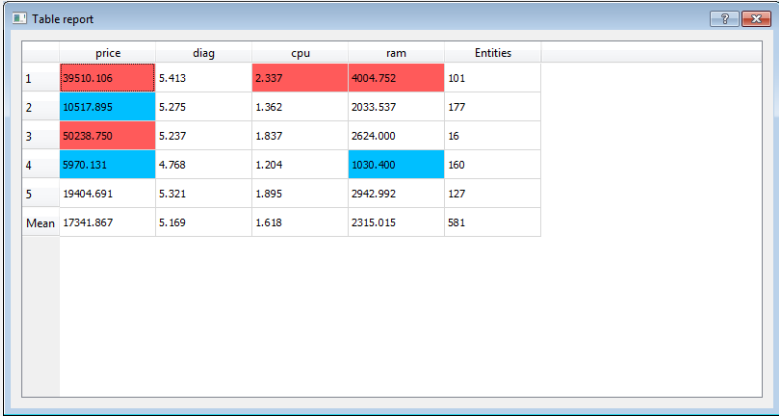


3. Посмотреть окно отчёта

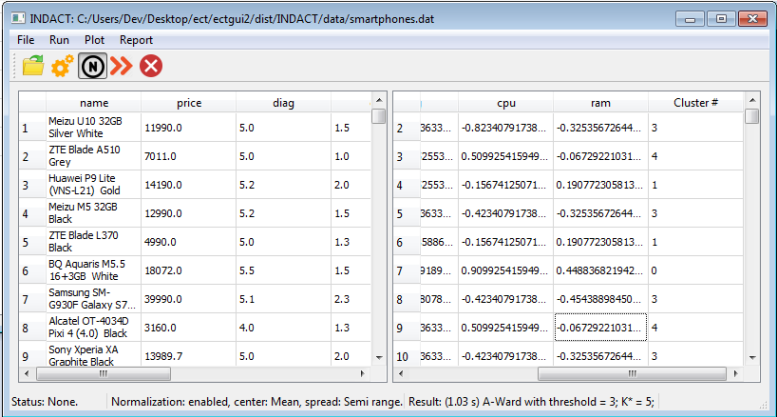
Окно текстового отчёта показано на рисунке справа и содержит только текстовое поле с табулированным отчётом. Из этого окна текст можно скопировать в файл.



5.8.2 Табличный отчёт

Действие/Описание	Интерфейс
<p>1. Сгенерировать отчёт</p> <p>Для генерации табличного отчёта в главном меню выбрать Report ⇒ Table.</p>	
<p>2. Посмотреть окно отчёта</p> <p>В окне табличного отчёта приведена сводная таблица в которой строки соответствуют кластерам, а столбцы — признакам. В ячейках указаны средние значения признака по кластеру. Красным цветом выделены ячейки, в которых относительная разность значения и средней величины признака по кластеру больше 30%, соответственно синим — меньше 30%. Маргинальная строка содержит средние значения признаков по всем кластерам, а столбец — число объектов в кластере.</p>	

5.9 Выход из программы

Действие/Описание	Интерфейс
<p>1. <i>Выйти из программы</i></p> <p>Для выхода из программы в главном меню выбрать File ⇒ Exit. Поздравляем, вы завершили работу с программой!</p>	

6 Алгоритмы кластеризации (краткое описание)

6.1 Алгоритм A-Ward

Алгоритм A-Ward является усовершенствованием широко известного алгоритма иерархической агломеративной кластеризации Уорда (Ward)[4]. На первом шаге все кластеры состоят из единственного объекта.

Остановка алгоритма происходит при достижении числа кластеров, заданного пользователем, или объединении всех объектов в едином универсальном кластере. Степень близости между двумя кластерами вычисляется как произведение квадрата евклидова расстояния между центрами кластеров и произведения численностей этих кластеров, делённого на их суммарную численность.

Недостаток алгоритма Уорда — медленность вычислений, связанная с необходимостью отыскания минимума расстояний, которых очень много на начальных этапах агломерации. В алгоритме A-Уорд эти шаги пропускаются, поскольку шаги агломерации применяются к некоторому предварительному разбиению объектов на достаточно малое число кластеров. Это-то предварительное разбиение используется как начальное для работы метода Уорда. Классы предварительного разбиения — это кластеры, полученные методом аномальной кластеризации.

Метод аномальной кластеризации находит и удаляет аномальные кластеры по одному до тех пор, пока не останется объектов для кластеризации. В основе этого метода лежит критерий квадратичной ошибки метода k -средних. Аномальным называется такой кластер, который наиболее удалён от начала координат, куда предварительно переносится центр данных. Его построение начинается с самого удалённого объекта, а затем в него добавляются все объекты, которые ближе к центру кластера, чем к точке начала отсчёта. Центр аномального кластера обновляется на каждом шаге, в то время как центр данных остаётся неизменным.

6.2 Алгоритм A-Ward _{$p\beta$}

Алгоритм A-Ward _{$p\beta$} — это дополнительная модификация для приложений, в которых требуется анализировать зашумленные данные, включающие нерелевантные признаки. В этом случае и Ward, и A-Ward плохо работают. Снизить влияние нерелевантных признаков позволяет введение весовых коэффициентов. В процессе работы алгоритма A-Ward _{$p\beta$} для каждого признака вычисляется вес, обратно пропорциональный его разбросу внутри кластера. При этом используется не обязательно евклидово расстояние, а метрика Минковского произвольной степени. Параметры p и β являются степенями Минковского и весовых коэффициентов признаков соответственно.

Как и в случае с A-Ward, алгоритм A-Ward _{$p\beta$} использует аномальную кластеризацию для предварительной “разведки” структуры данных и снижения времени работы, однако в алгоритме A-Ward _{$p\beta$} аномальная кластеризация обобщена с учётом дополнительных параметров.

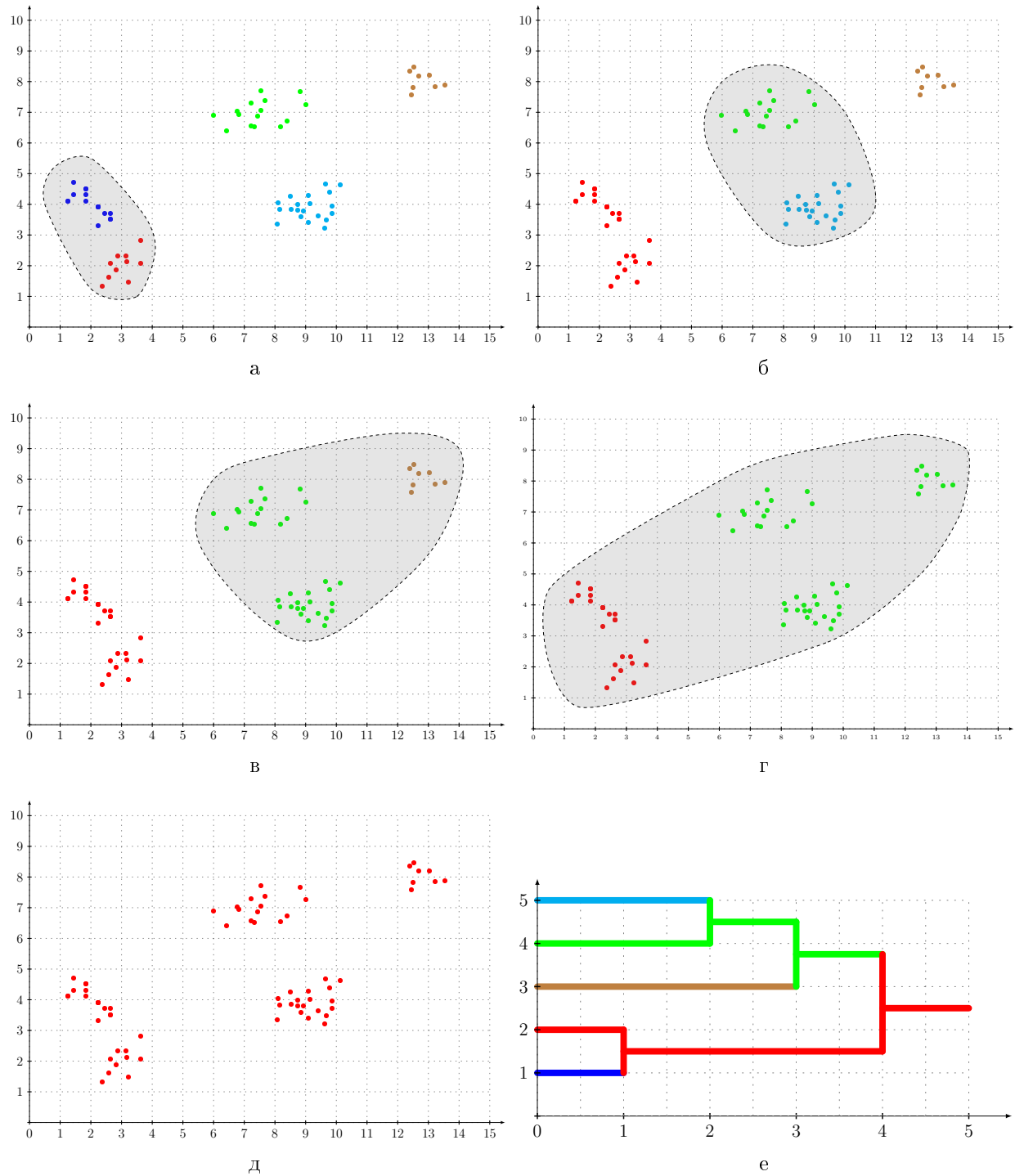


Рисунок 11 – Принцип работы A-Ward

6.3 Алгоритм BiKM-R

Алгоритм BiKM-R (bisecting k-means randomized / Раздвоение по методу k-средних) относится к классу дивизивных алгоритмов иерархического кластер-анализа. В отличие от агломеративных алгоритмов, где вычисления организованы “снизу-вверх” путём объединения, здесь вычисления организованы “сверху-вниз” путём разделения кластеров, начиная с универсального кластера, состоящего из всех объектов. На каждом шаге определённый кластер S разбивается на два по критерию суммы квадратов ошибок. Для инициализации алгоритма требуется указать начальные центры c_1 и c_2 . Затем осуществляются двухшаговые итерации по методу k-средних при $k = 2$. На первом шаге обновляются кластеры, путём разделения объектов на тех, что ближе к c_1 (кластер S_1) и тех, что ближе к c_2 (кластер S_2). На втором шаге вычисляются новые центры S_1 и S_2 . Процесс заканчивается, как только новые центры совпадают со старыми. Как и в случае с агломеративным алгоритмом, выбор c_1 и c_2 может быть организован с использованием метода аномальных кластеров. Для инициализации алгоритма раздвоения используются центры двух наибольших аномальных кластеров.

Для остановки алгоритма BiKM-R используется критерий, основанный на проецировании точек кластеров на случайные направления. Пусть на некотором этапе работы алгоритма имеется K кластеров. Генерируются s случайных векторов p_i , $i = 1, \dots, s$. Для генерации используется нормальное сферическое распределение со средним в начале координат и $\sigma^2 = 1/V$, где V – количество признаков. Затем каждый элемент x каждого кластера S_k ($k = 1, \dots, K$) проецируется на направления p_i , координаты проекции определяются как скалярное произведение: $x_i = \langle x, p_i \rangle$. Для каждого направления вычисляется функция плотности f_k^i по методу ядерной оценки (окно Парзена). Если для некоторого кластера S_k отношение ϵ_k числа направлений, для которых функции плотности f_k^i имеют по крайней мере один минимум, к общему числу направлений меньше заданного пользователем порога ϵ , то кластер S_k не разбивается. Для разделения выбирается в первую очередь кластер с наибольшим отношением ϵ_k/ϵ . Выбранный кластер разбивается по наиболее глубокому минимуму функции плотности.

6.4 Алгоритм dePDDP

Алгоритм dePDDP (Principal Direction Divisive Partitioning) относится к иерархическим дивизивным. Первоначально критерий разделения кластера на две части был относительно простым: предлагалось разделить кластер по его главной компоненте на положительную и отрицательную части. В алгоритме dePDDP эта идея усовершенствована при помощи правила, учитывающего распределение данных. Разбиение производится по наиболее глубокому минимуму функции плотности данных, спроецированных на первую главную компоненту данного кластера. Это правило используется для решения двух сопряжённых проблем: выбора кластера для разбиения и остановки алгоритма. Для разбиения выбирается кластер с наименьшим минимумом среди всех терминальных кластеров. Если кластер имеет монотонную или выпуклую функцию

плотности, то такой кластер не может быть разделен по критерию данного алгоритма. Экспериментально было показано, что алгоритм, работающий на описанных принципах эффективно решает задачу кластеризации как на реальных данных, так и на синтетических. Оценка функции плотности осуществляется по методу ядерной оценки (окно Парзена).

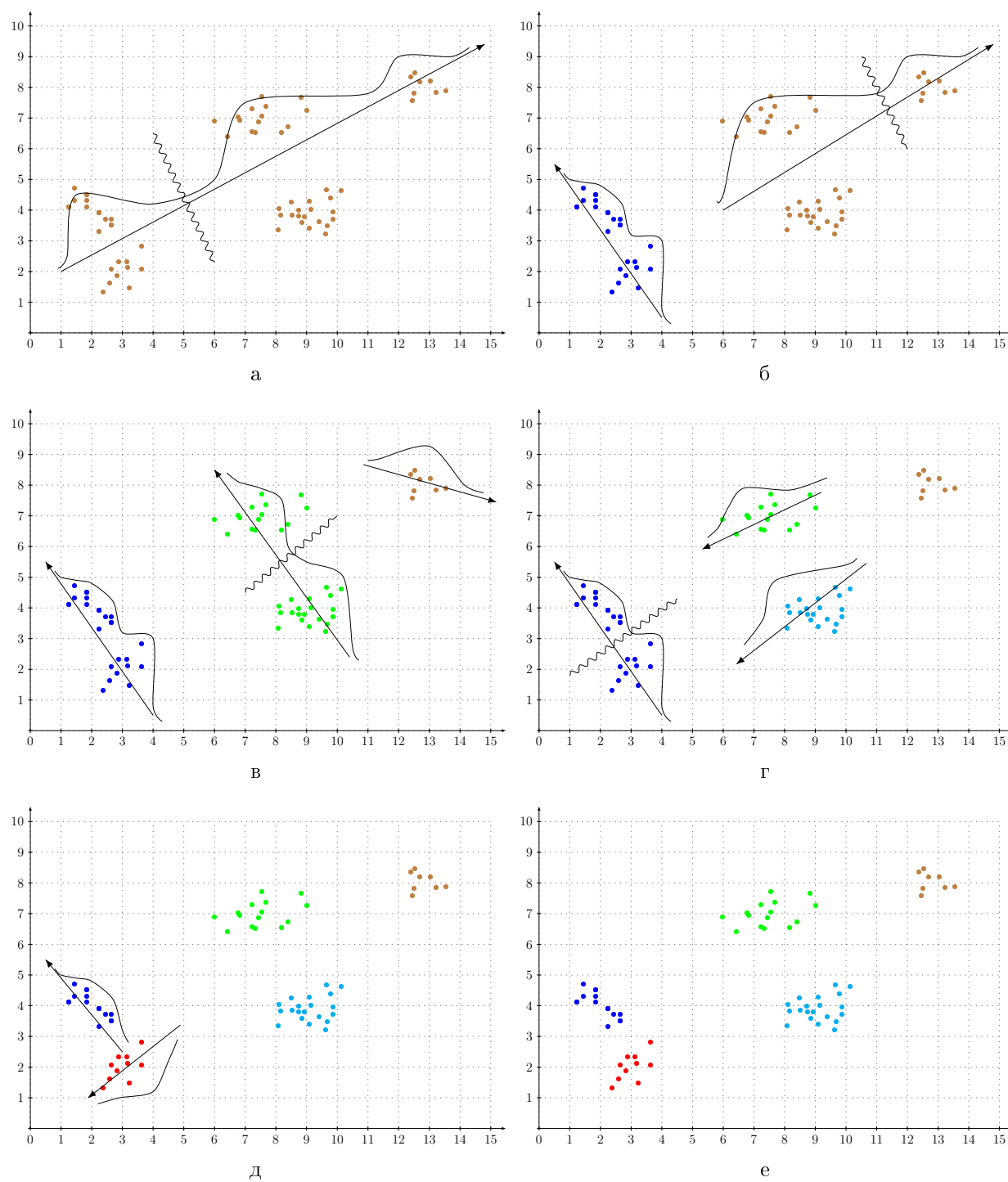


Рисунок 12 – Принцип работы dePDDP

7 Демонстрация работы программы

7.1 Нормализация

7.1.1 Нормализация с центрированием по среднему и масштабированием по полуразмаху

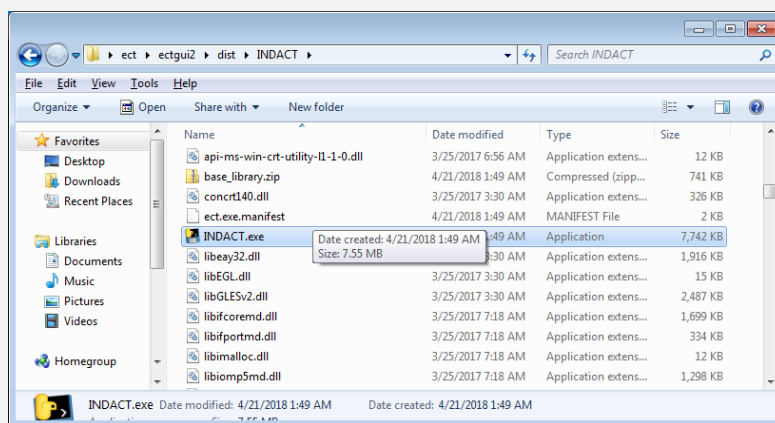
В данном разделе рассматривается пример нормализации признаков обучающего файла `smartphones.dat` с центрированием по среднему и масштабированием по полуразмаху.

Действие/Описание

Интерфейс

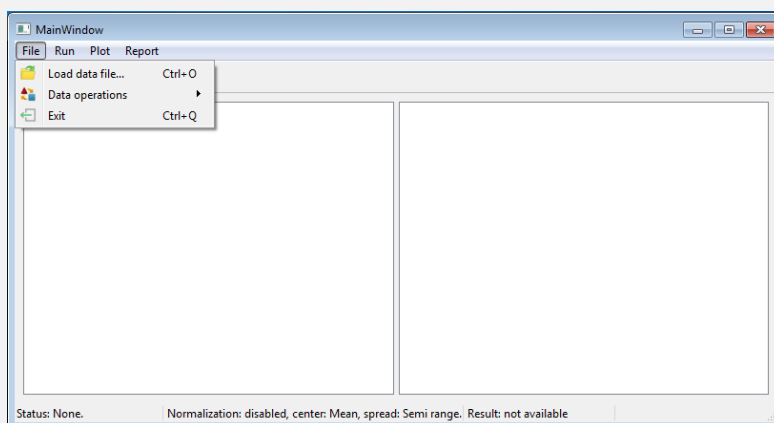
1. Запустить бинарный файл программы

Дважды нажать левой кнопкой мыши (ЛКМ) на значке `INDACT.exe`.



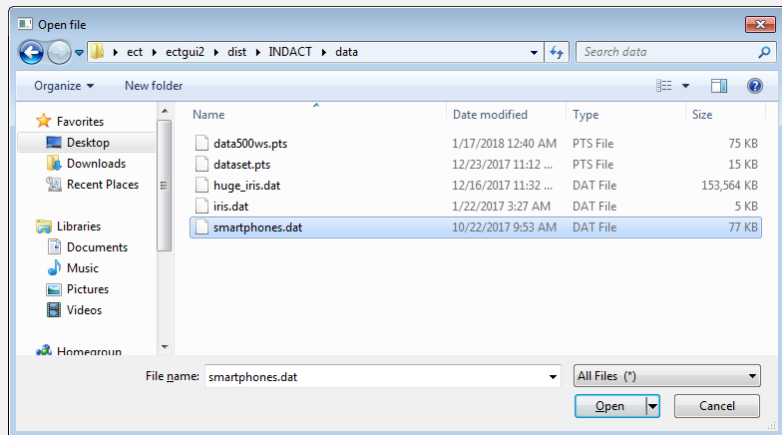
2. Открыть диалог загрузки файла

Последовательно нажать в главном меню пункты **File** ⇒ **Load data file**.



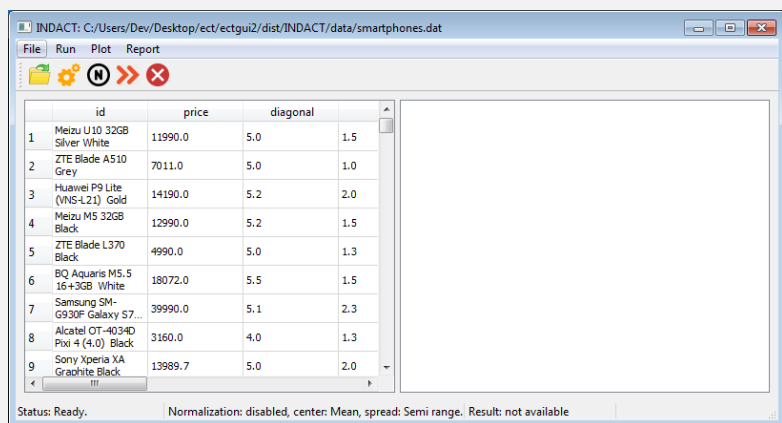
3. Выбрать текстовый файл с данными

В файловом диалоге выбрать загружаемый файл `INDACT/data/smartphones.dat` и нажать кнопку **Open**.



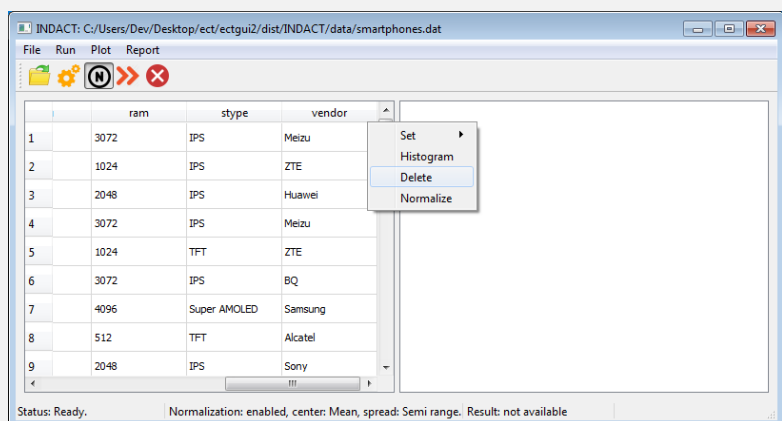
4. Убедиться что файл загружен

Посмотреть, что панель исходных данных заполнилась данными из файла.



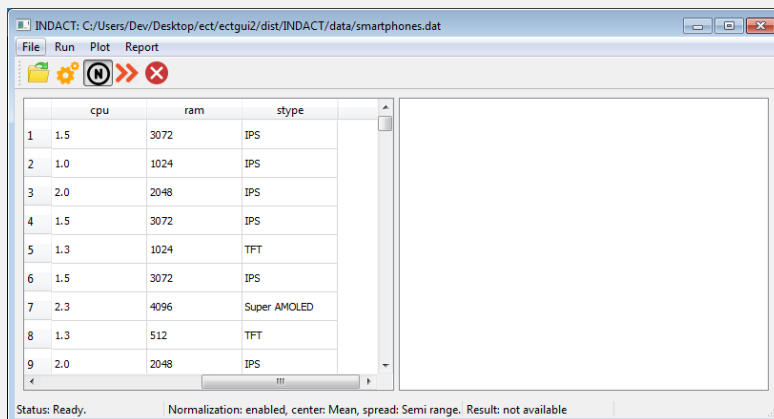
5. Удалить признак vendor

Вызвать контекстное меню на признаке `vendor` при помощи ПКМ и выбрать пункт **Delete**.



6. Убедиться, что признак удалился


Результат удаления признака **vendor** показан на рисунке справа. Видно, что признак **vendor** больше не отображается на панели исходных данных.

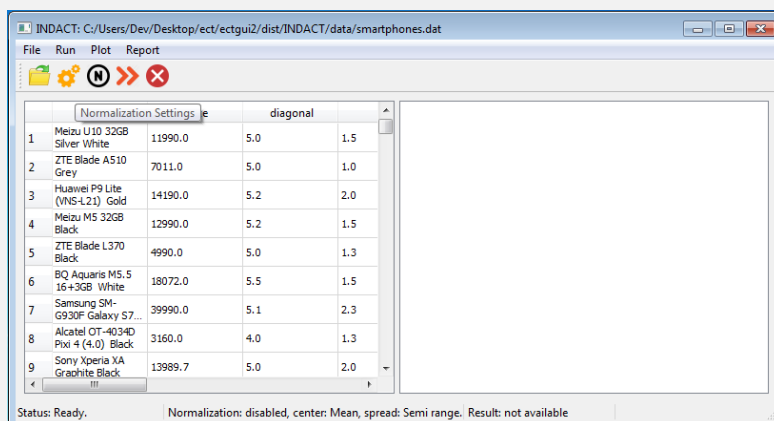


	cpu	ram	stype
1	1.5	3072	IPS
2	1.0	1024	IPS
3	2.0	2048	IPS
4	1.5	3072	IPS
5	1.3	1024	TFT
6	1.5	3072	IPS
7	2.3	4096	Super AMOLED
8	1.3	512	TFT
9	2.0	2048	IPS

Status: Ready. Normalization: enabled, center: Mean, spread: Semi range. Result: not available

7. Открыть окно нормализации

Диалог нормализации можно открыть щелчком ЛКМ на иконке  (см. рисунок 1 обозначение 11)

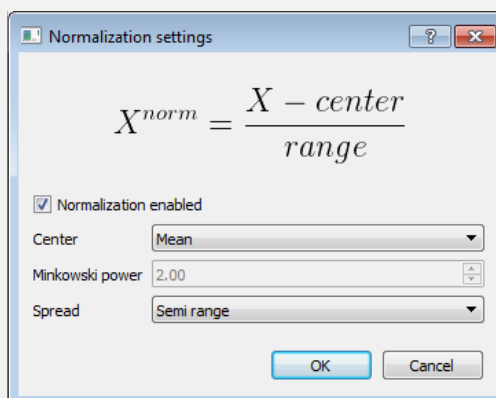


	diagonal	le
1	Meizu U10 32GB Silver White	11990.0
2	ZTE Blade A510 Grey	7011.0
3	Huawei P9 Lite (VNS-L21) Gold	14190.0
4	Meizu M5 32GB Black	12990.0
5	ZTE Blade L370 Black	4990.0
6	BQ Aquaris M5.5 16+3GB White	18072.0
7	Samsung SM-G930F Galaxy S7...	39990.0
8	Alcatel OT-4034D Pixi 4 (4.0) Black	3160.0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

Status: Ready. Normalization: disabled, center: Mean, spread: Semi range. Result: not available

8. Выставить параметры

Выставить параметры нормализации, как показано на рисунке справа. Переключатель “Normalization enabled” должен быть включен, значение **Center** выбрано **Mean**, а значение **Range** — **Semi range**. Подтвердить ввод **OK**.




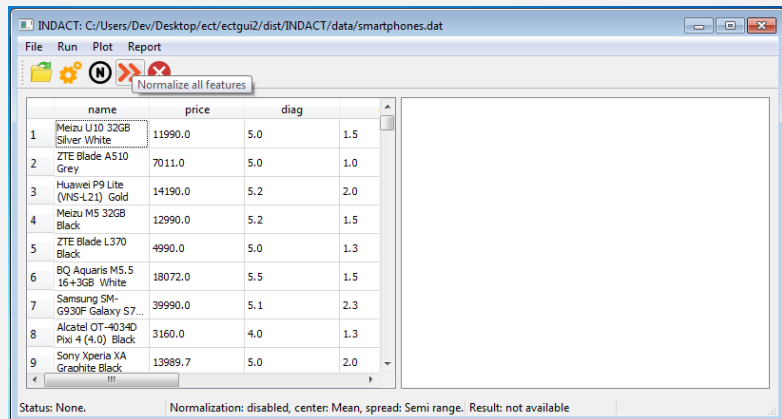
$$X^{norm} = \frac{X - center}{range}$$

☒ Normalization enabled
 Center: **Mean**
 Minkowski power: **2.00**
 Spread: **Semi range**

OK Cancel

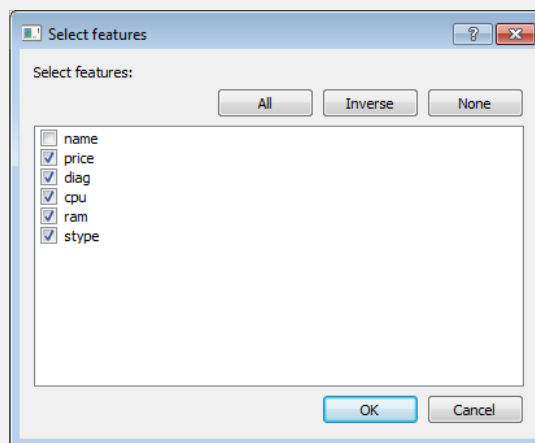
9. Запустить нормализацию нескольких признаков

Чтобы нормализовать несколько признаков сразу, требуется нажать иконку . Откроется окно выбора признаков.



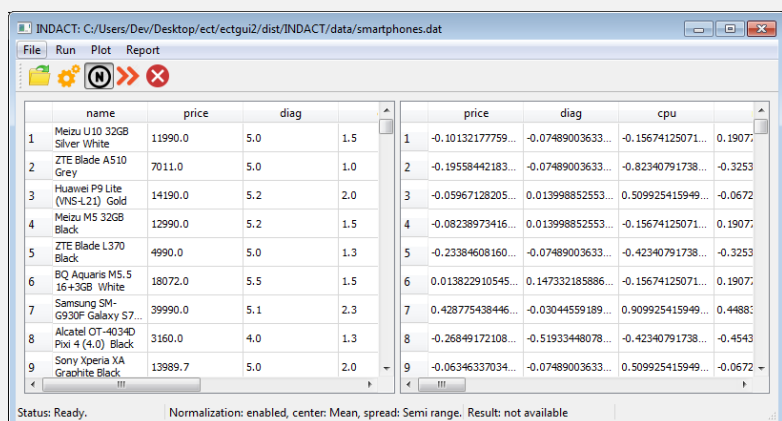
10. Выбрать признаки для нормализации

В открытом окне выбрать все признаки, за исключением признака `name`. Для этого нажать кнопку **All** и после снять выделение с `name`. Подтвердить выбор кнопкой **Ok**.



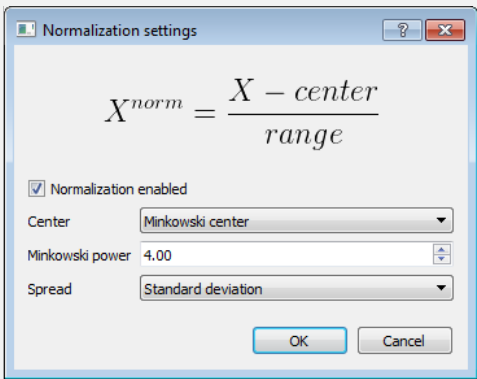
11. Проверить результат нормализации

После выполнения нормализации выбранные признаки будут добавлены на панель нормализованных данных, при этом категориальный признак `stype` разложится в бинарный.



7.1.2 Нормализация с центрированием по Минковскому и масштабированием по стандартному отклонению

Теперь рассмотрим пример нормализации данных из демонстрационного примера с центрированием Минковского.

Действие/Описание	Интерфейс
1. Запустить программу и загрузить файл	Выполнить пункты 1–7 из предыдущего примера (7.1.1).
<p>2. Выставить параметры нормализации</p> <p>Выставить параметры нормализации, как показано на рисунке справа. Переключатель “Normalization enabled” должен быть включён, значение поля Center — Minkowski Center, величина Minkowski Power выставлена равной 4, а параметр Range — Standard deviation.</p>	
3. Нормализовать все признаки	Выполнить пункты 9,10 из предыдущего примера (7.1.1).

4. Посмотреть результат нормализации

После выполнения нормализации выбранные признаки будут добавлены на панель нормализованных данных.

	cpu	ram	stype
1	1.5	3072	IPS
2	1.0	1024	IPS
3	2.0	2048	IPS
4	1.5	3072	IPS
5	1.3	1024	TFT
6	1.5	3072	IPS
7	2.3	4096	Super AMOLED
8	1.3	512	TFT
9	2.0	2048	IPS

	price	diag	cpu
1	-0.90756380546...	-0.08212737419...	-0.44211665303...
2	-1.24028602890...	-0.08212737419...	-1.59159520265...
3	-0.76054856313...	0.336926860291...	0.707361896589...
4	-0.84073869531...	0.336926860291...	-0.44211665303...
5	-1.37533957651...	-0.08212737419...	-0.90190807288...
6	-0.50113348553...	0.965508212027...	-0.44211665303...
7	0.963539278743...	0.127399743045...	1.397049026362...
8	-1.49762952809...	-2.17739854665...	-0.90190807288...
9	-0.77393363269...	-0.08212737419...	0.707361896589...

7.2 Кластеризация

7.2.1 Кластеризация с автоматическим выбором числа кластеров

Рассмотрим пример кластеризации с использованием метода, который предусматривает автоматическое вычисление числа кластеров в процессе работы. Используем для этого процедуру нормализации, проиллюстрированную ранее. Выберем типичные значения параметров нормализации: центрирование по среднему, масштабирование полуразмахом (см. 7.1.1).

Действие/Описание

Интерфейс

1. Запустить программу, загрузить файл и нормализовать признаки

Выполнить все пункты из первого примера(7.1.1). Для кластеризации требуются нормализованные признаки.

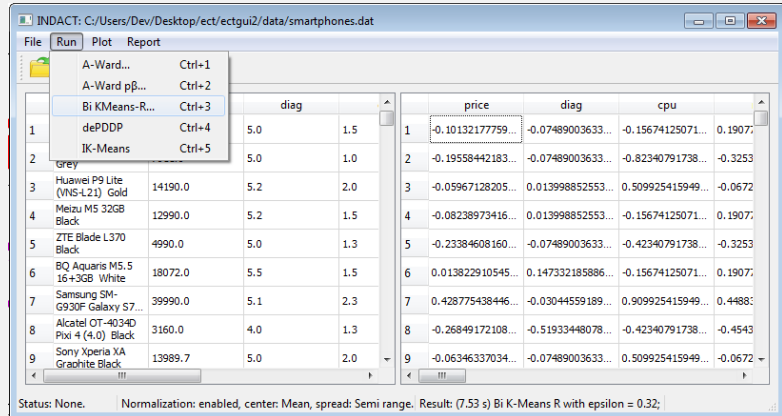
	name	price	diag
1	Meizu U10 32GB Silver White	11990.0	5.0
2	ZTE Blade A510 Grey	7011.0	5.0
3	Huawei P9 Lite (VNS-L21) Gold	14190.0	5.2
4	Meizu M5 32GB Black	12990.0	5.2
5	ZTE Blade L370 Black	4990.0	5.0
6	BQ Aquaris M5.5 16+3GB White	18072.0	5.5
7	Samsung SM-G930F Galaxy S7...	39990.0	5.1
8	Alcatel OT-4034D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

	price	diag	cpu
1	-0.10132177759...	-0.07489003633...	-0.15674125071...
2	-0.19558442183...	-0.07489003633...	-0.82340791738...
3	-0.05967128205...	0.013998852553...	0.509925415949...
4	-0.08238973416...	0.013998852553...	-0.15674125071...
5	-0.23384608160...	-0.07489003633...	-0.42340791738...
6	0.013822910545...	0.147332185886...	-0.15674125071...
7	0.428775438446...	-0.03044559189...	0.909925415949...
8	-0.26849172108...	-0.51933448078...	-0.42340791738...
9	-0.06346337034...	-0.07489003633...	0.509925415949...

2. Открыть окно кластеризации

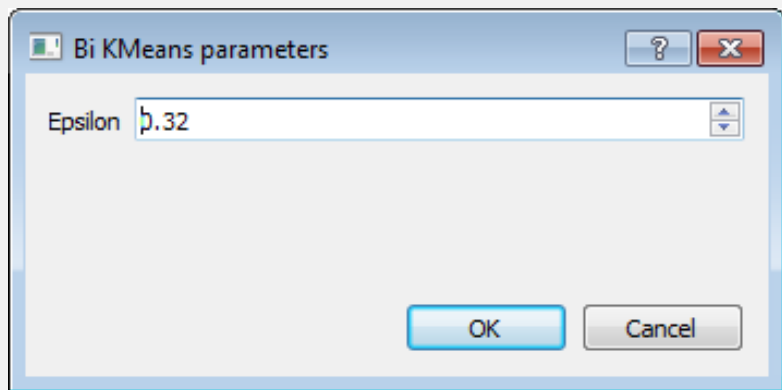
Для запуска кластеризации требуется открыть окно с настройками кластеризации. В данном примере рассматривается автоматическое определение числа кластеров, поэтому в главном меню выберем **Run** ⇒

Bi KMeans-R.



3. Задать параметр алгоритма

Алгоритм Bi K-Means R требует задания единственного параметра ϵ . В данном случае рекомендуется оставить значение по умолчанию. Подтвердить ввод кнопкой **Ok**



4. Дождатся завершения работы алгоритма

Выполнение алгоритма требует некоторого времени, после завершения работы краткая информация о последнем запущенном алгоритме и времени работы будет отображена в строке состояния. Например, на рисунке справа видно, что алгоритм Bi K-Means R при $\epsilon = 0.32$ завершился за 6,23 сек.

INDACT: C:/Users/Dev/Desktop/ectgui2/data/smartphones.dat

	name	price	diag		ZO]	stype[OLED]	stype[LCD]	Cluster # (C)
1	Meizu U10 32GB Silver White	11990.0	5.0	1.5	1	3855...	-0.00092000427...	2
2	ZTE Blade A510 Grey	7011.0	5.0	1.0	2	3855...	-0.00092000427...	5
3	Huawei P9 Lite (VNS-L21) Gold	14190.0	5.2	2.0	3	3855...	-0.00092000427...	0
4	Meizu M5 32GB Black	12990.0	5.2	1.5	4	3855...	-0.00092000427...	2
5	ZTE Blade L370 Black	4990.0	5.0	1.3	5	3855...	-0.00092000427...	1
6	BQ Aquaris M5.5 16+3GB White	18072.0	5.5	1.5	6	3855...	-0.00092000427...	2
7	Samsung SM-G930F Galaxy S7...	39990.0	5.1	2.3	7	3855...	-0.00092000427...	0
8	Alcatel OT-4034D Pixi 4 (4.0) Black	3160.0	4.0	1.3	8	3855...	-0.00092000427...	1
9	Sony Xperia XA Gracihite Black	13989.7	5.0	2.0	9	3855...	-0.00092000427...	0

Status: None. Normalization: enabled, center: Mean, spread: Semi range. Result: (6.23 s) Bi K-Means R with epsilon = 0.32;

5. Посмотреть кластеры

После завершения алгоритма в столбце Cluster # будут показаны метки кластерной принадлежности объектов, как показано на рисунке справа. Следует иметь ввиду, что алгоритм Bi K-Means R использует в работе случайно сгенерированные направления, поэтому для различных запусков при разной инициализации генератора случайных чисел результаты могут отличаться.

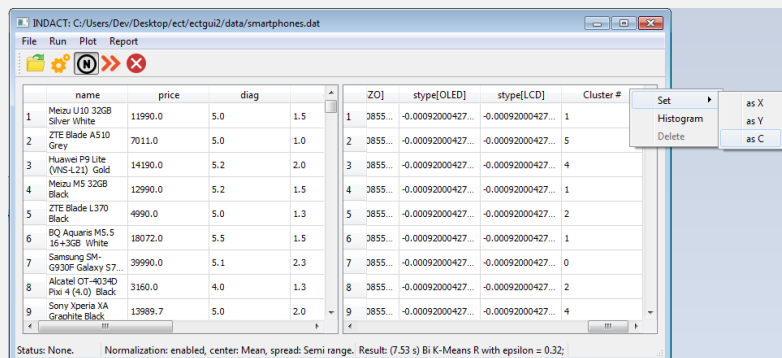
INDACT: C:/Users/Dev/Desktop/ectgui2/data/smartphones.dat

	name	price	diag		ZO]	stype[OLED]	stype[LCD]	Cluster # (C)
1	Meizu U10 32GB Silver White	11990.0	5.0	1.5	1	3855...	-0.00092000427...	2
2	ZTE Blade A510 Grey	7011.0	5.0	1.0	2	3855...	-0.00092000427...	5
3	Huawei P9 Lite (VNS-L21) Gold	14190.0	5.2	2.0	3	3855...	-0.00092000427...	0
4	Meizu M5 32GB Black	12990.0	5.2	1.5	4	3855...	-0.00092000427...	2
5	ZTE Blade L370 Black	4990.0	5.0	1.3	5	3855...	-0.00092000427...	1
6	BQ Aquaris M5.5 16+3GB White	18072.0	5.5	1.5	6	3855...	-0.00092000427...	2
7	Samsung SM-G930F Galaxy S7...	39990.0	5.1	2.3	7	3855...	-0.00092000427...	0
8	Alcatel OT-4034D Pixi 4 (4.0) Black	3160.0	4.0	1.3	8	3855...	-0.00092000427...	1
9	Sony Xperia XA Gracihite Black	13989.7	5.0	2.0	9	3855...	-0.00092000427...	0

Status: None. Normalization: enabled, center: Mean, spread: Semi range. Result: (6.23 s) Bi K-Means R with epsilon = 0.32;

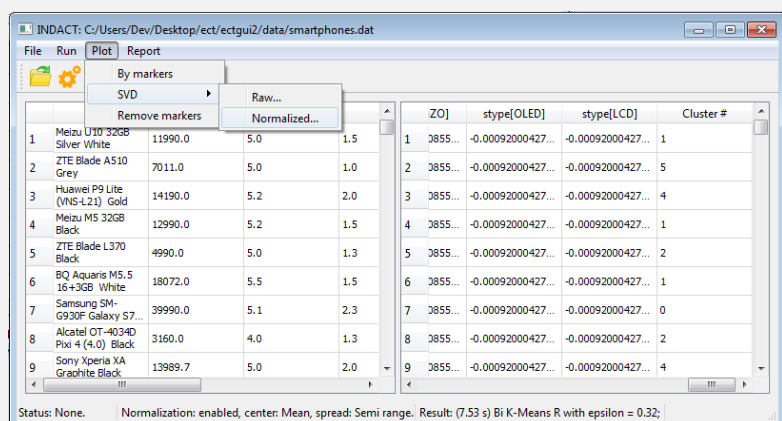
6. Установить метку C

Для того чтобы обозначить, что полученное разбиение будет теперь определять цвет на диаграммах, требуется выставить метку C на столбец Cluster #. Для этого в контекстном меню выбрать соответствующие пункты, как показано на рисунке справа.



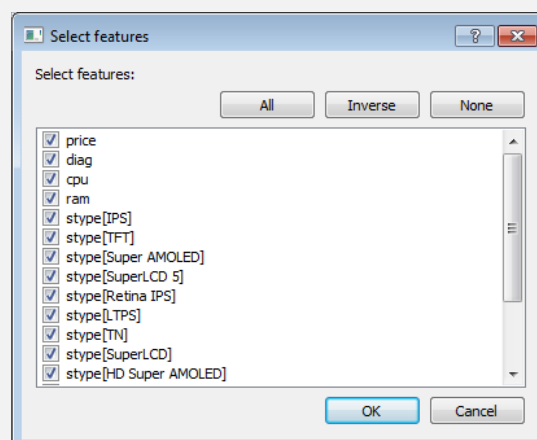
7. Построить диаграмму SVD

Для общей оценки полученного разбиения можно построить SVD диаграмму. Для этого в главном меню выбрать **Plot** ⇒ **SVD** ⇒ **Normalized**.



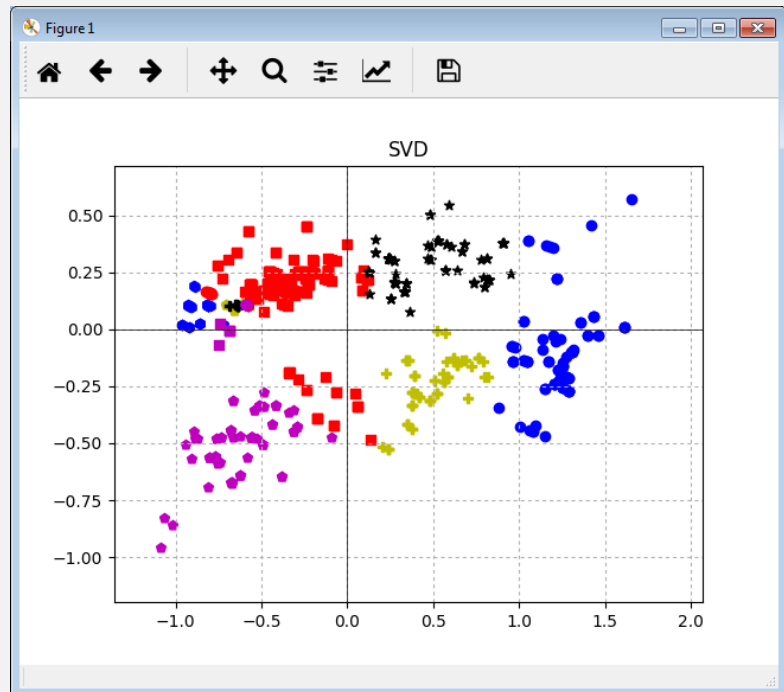
8. Выбрать признаки

Для построения диаграммы выбрать все признаки, которые участвовали в кластеризации.



9. Посмотреть SVD диаграмму

Полученная диаграмма показана на рисунке справа.



10. Сгенерировать текстовый отчёт

Для генерации текстового отчёта в главном меню выбрать пункты **Report** ⇒ **Text**.

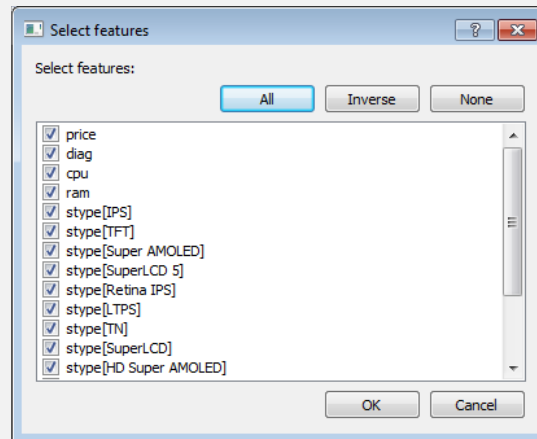
The screenshot shows the INDACT software interface. The main window displays a table of smartphone data. A 'Report' window is open, showing a 'Text' report. The table has columns for name, price, diag, and other attributes. The report window shows a list of data points with their corresponding cluster numbers.

name	price	diag	Cluster # (C)
1 Meizu U10 32GB Silver White	11990.0	5.0	1.5
2 ZTE Blade A510 Grey	7011.0	5.0	1.0
3 Huawei P9 Lite (WNS-L21) Gold	14190.0	5.2	2.0
4 Meizu M5 32GB Black	12990.0	5.2	1.5
5 ZTE Blade L370 Black	4990.0	5.0	1.3
6 BQ Aquaris M5.5 16+3GB White	18072.0	5.5	1.5
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 Gracihite Black	13989.7	5.0	2.0

Status: None. Normalization: enabled, center: Mean, spread: Semi range. Result: (6.23 s) Bi K-Means R with epsilon = 0.32;

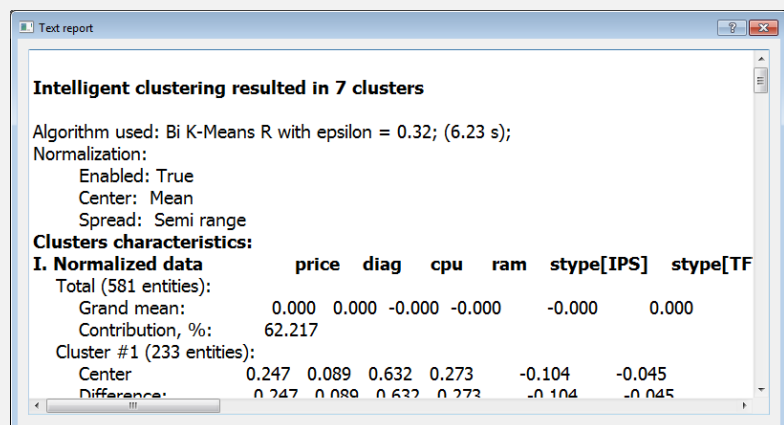
11. Выбрать признаки

В открывшемся окне выбрать признаки, по которым будут включены в отчёт. Для быстрого выбора всех признаков, можно нажать кнопку **All**. Подтвердить выбор кнопкой **Ok**.



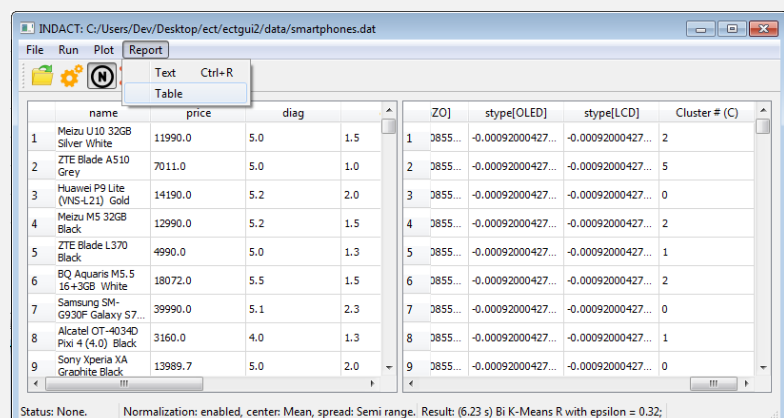
12. Посмотреть текстовый отчёт

Посмотреть текстовый отчёт в открывшемся окне. Вид окна показан на рисунке справа.



13. Сгенерировать табличный отчёт

Для генерации текстового отчёта в главном меню выбрать пункты **Report** ⇒ **Table**.



*14. Посмотреть
табличный отчёт*

Посмотреть табличный от-
чет в открывшемся окне.

	price	diag	cpu	ram	stype[IPS]	stype[TFT]	'pe[Super AMOLED	'type[S
1	30390.242	5.370	2.092	3397.219	0.412	0.000	0.219	0.030
2	6327.033	4.614	1.246	1108.164	0.000	0.630	0.096	0.000
3	13580.616	5.116	1.518	2649.825	0.702	0.053	0.123	0.000
4	9562.489	5.354	1.325	1824.000	0.982	0.000	0.000	0.000
5	5884.606	4.955	1.287	985.212	1.000	0.000	0.000	0.000
6	6841.097	4.886	1.055	1024.000	1.000	0.000	0.000	0.000
7	9177.745	5.000	1.000	2048.000	1.000	0.000	0.000	0.000
Mean	17341.867	5.169	1.618	2315.015	0.606	0.084	0.112	0.012

7.2.2 Кластеризация с заданным числом кластеров

Если известно конкретное число кластеров входящих в состав данных, можно применить методы, подразумевающие явный ввод с клавиатуры. Метод A-Ward позволяет задать число кластеров (см. раздел 6.1).

Действие/Описание

1. Запустить программу, загрузить файл и нормализовать признаки

Выполнить все пункты из второго примера по нормализации (7.1.2). Для кластеризации требуются нормализованные признаки. Для данного примера используются нормализация с центрированием Минковского.

Интерфейс

	cpu	ram	stype
1	1.5	3072	IPS
2	1.0	1024	IPS
3	2.0	2048	IPS
4	1.5	3072	IPS
5	1.3	1024	TFT
6	1.5	3072	IPS
7	2.3	4096	Super AMOLED
8	1.3	512	TFT
9	2.0	2048	IPS

	price	diag	cpu
1	-0.90756380546...	-0.08212737419...	-0.44211665303...
2	-1.24028602890...	-0.08212737419...	-1.59159520265...
3	-0.76054856313...	0.336926860291...	0.707361896589...
4	-0.84073869531...	0.336926860291...	-0.44211665303...
5	-1.37533957651...	-0.08212737419...	-0.90190807288...
6	-0.50113348553...	0.965508212027...	-0.44211665303...
7	0.963539278743...	0.127399743045...	1.397049026362...
8	-1.49762952809...	-2.17739854665...	-0.90190807288...
9	-0.77393363269...	-0.08212737419...	0.707361896589...

2. Выбрать алгоритм A-Ward_{pβ}

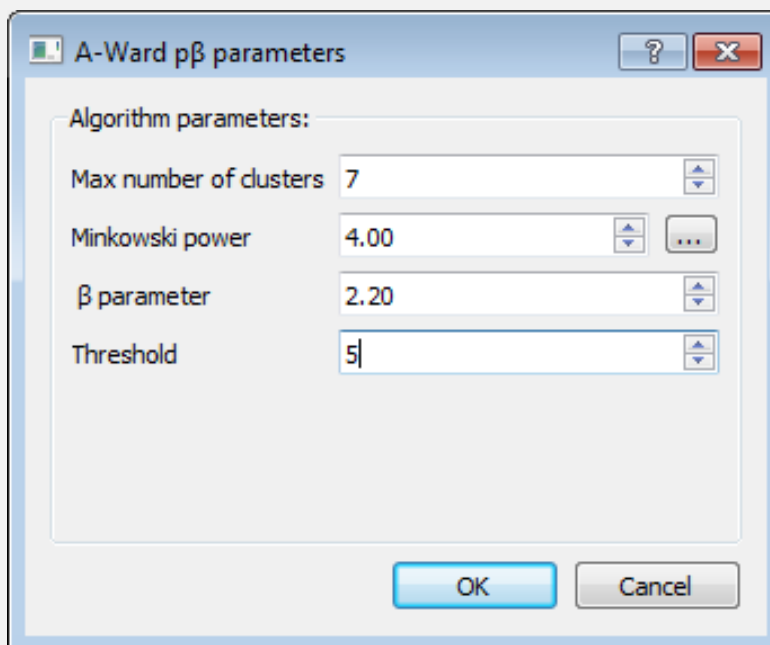
В главном меню выбрать пункт **Run** ⇒ **A-Ward_{pβ}**

	cpu	ram	stype
1	1.5	3072	IPS
2	1.0	1024	IPS
3	2.0	2048	IPS
4	1.5	3072	IPS
5	1.3	1024	TFT
6	1.5	3072	IPS
7	2.3	4096	Super AMOLED
8	1.3	512	TFT
9	2.0	2048	IPS

	price	diag	cpu
1	-0.9076000559...	-0.08212846292...	-0.44211793749...
2	-1.24032222903...	-0.08212846292...	-1.59159648711...
3	-0.76058476326...	0.336925771563...	0.707360612125...
4	-0.84077489544...	0.336925771563...	-0.44211793749...
5	-1.37537577665...	-0.08212846292...	-0.90190935734...
6	-0.50116968566...	0.965507123299...	-0.44211793749...
7	0.9635303078609...	0.127398654317...	1.397047741899...
8	-1.49766572822...	-2.17739963538...	-0.90190935734...
9	-0.77396983283...	-0.08212846292...	0.707360612125...

3. Задать настройки

В открывшемся диалоговом окне задать настройки алгоритма, как показано на рисунке справа. Подтвердить ввод кнопкой **Ok**.



4. Дождаться результата кластеризации

Выполнение алгоритма тербует некоторого времени, после завершения работы краткая информация о последнем запущенном алгоритме и времени работы будет отображена в строке состояния.

Например, на рисунке справа видно, что алгоритм $A-Ward_{p\beta}$ при заданных параметрах завершился за 0,93 сек.

INDACT: C:/Users/Dev/Desktop/ect/ectgui2/data/smartphones.dat

File Run Plot Report

	name	price	diag			ZO]	stype[OLED]	stype[LCD]	Cluster #	
1	Meizu U10 32GB Silver White	11990.0	5.0	1.5		1	0667...	-0.68975914702...	-0.68975826796...	4
2	ZTE Blade A510 Grey	7011.0	5.0	1.0		2	0667...	-0.68975914702...	-0.68975826796...	4
3	Huawei P9 Lite (VNS-L21) Gold	14190.0	5.2	2.0		3	0667...	-0.68975914702...	-0.68975826796...	4
4	Meizu M5 32GB Black	12990.0	5.2	1.5		4	0667...	-0.68975914702...	-0.68975826796...	4
5	ZTE Blade L370 Black	4990.0	5.0	1.3		5	0667...	-0.68975914702...	-0.68975826796...	4
6	BQ Aquaris M5.5 16+3GB White	18072.0	5.5	1.5		6	0667...	-0.68975914702...	-0.68975826796...	4
7	Samsung SM-G930F Galaxy S7...	39990.0	5.1	2.3		7	0667...	-0.68975914702...	-0.68975826796...	4
8	Alcatel OT-4034D Pixi 4 (4.0) Black	3160.0	4.0	1.3		8	0667...	-0.68975914702...	-0.68975826796...	4
9	Sony Xperia XA Graphite Black	13989.7	5.0	2.0		9	0667...	-0.68975914702...	-0.68975826796...	4

St Normalization: enabled, center: Minkowski center, spread: Standard deviation, mi: Result: (0.934 s) A-Ward_p_beta with threshold = 5; p = 4.0; K*

5. Построить диаграммы и отчёты

Повторить операции 5–14 из предыдущего примера [7.2.1](#)

Список литературы

- [1] de Amorim R.C. Mirkin B. Minkowski metric, feature weighting and anomalous cluster initializing in k-means clustering // Pattern Recognition. 2012. № 03. С. 1061–1075.
- [2] Миркин Б. Г. Введение в анализ данных. М.: Юрайт, 2015.
- [3] Kovaleva E.V. Mirkin B.G. Bisecting K-Means and 1D Projection Divisive Clustering: A Unified Framework and Experimental Comparison // Journal of Classification. 2015. № 10. С. 414–444.
- [4] Joe H. Ward J. Hierarchical Grouping to Optimize an Objective Function // Journal of American Statistical Association. 1963.
- [5] Boley D. Principal Direction Divisive Partitioning // Data Mining and Knowledge Discovery. 1998. № 02. С. 325–344.
- [6] Tasoulis S.K. Tasoulis D.K. Plagianakos V.P. Enhancing Principal Direction Divisive Clustering // Pattern Recognition. 2010. № 43. С. 3391–3411.
- [7] Mirkin B. Core Concepts in Data Analysis: Summarization, Correlation, Visualization.