

20/04/2018 v3.0
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. Для практического применения этот недостаток зачастую вынуждает отказаться от использования k -средних. В этом случае задачу позволяют решить интеллектуальные методы, которые в процессе работы или другими способами позволяют автоматически определить число кластеров. Программная система INDACT обладает всем необходимым функционалом и включает в свой инструментарий множество методов, необходимых для решения сложных задач кластер анализа.

Содержание

1	Введение	4
1.1	Область применения	4
1.2	Описание возможностей	4
1.3	Уровень подготовки пользователя	4
1.4	Исходный код	4
2	Назначение	5
3	Условия применения и подготовка к работе	6
4	Основные принципы работы	7
4.1	Этапы работы с программой	7
4.2	Требования к файлу исходных данных	7
4.3	Обучающий файл	8
4.4	Нормализация	8
4.5	Просмотр результатов кластеризации	9
4.6	Общие сведения о пользовательском интерфейсе	10
4.6.1	Главное окно	10
4.6.2	Контекстное меню	12
4.6.3	Диалог нормализации	13
4.6.4	Окно графического вывода	14
4.6.5	Окно генерации данных	15
4.6.6	Окно запуска кластеризации	16
4.6.7	Окно таблицы результатов	18
5	Описание операций	19
5.1	Запуск программы	19
5.2	Загрузка исходных данных	20
5.3	Нормализация	21
5.3.1	Установка параметров нормализации	21
5.3.2	Нормализация одного признака	23
5.3.3	Нормализация всех признаков сразу	26
5.4	Отбор признаков	28
5.4.1	Удаление одного признака	28
5.4.2	Удаление всех признаков сразу	30
5.4.3	Установка признака как индекс	30
5.5	Настройка способа отображения вкладок	33
5.6	Визуализация	34

5.6.1	Построение гистограммы по признаку	34
5.6.2	Построение поля рассеяния (scatter plot)	35
5.6.3	Построение SVD диаграммы	37
5.7	Генерация синтетических данных	38
5.8	Запуск кластеризации	41
5.9	Генерация отчёта	42
5.10	Выход из программы	46
6	Алгоритмы кластеризации (краткое описание)	47
6.1	Алгоритм $A - Ward$	47
6.2	Алгоритм $A - Ward_{p\beta}$	47
6.3	Алгоритм BiKM-R	48
6.4	Алгоритм dePDDP	48
7	Примеры работы с программой	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	55
7.2	Кластеризация	57
7.2.1	Кластеризация с автоматическим выбором числа кластеров	57
7.2.2	Кластеризация с явно заданным числом кластеров	61
	Аббревиатуры	64
	Словарь терминов	66
	Список литературы	69

1 Введение

1.1 Область применения

Программное обеспечение INDACT применяется для проведения кластер-анализа таблиц данных с использованием интеллектуальных алгоритмов. Типичный пример задачи для решения которой может применяться кластер-анализ — задача об ирисах Фишера. Эта задача состоит в поиске 50 экземпляров каждого из трёх видов ириса — Ирис щетинистый (*Iris setosa*), Ирис виргинский (*Iris virginica*) и Ирис разноцветный (*Iris versicolor*) на данных из 150 объектов. Каждый объект обладает четырьмя признаками:

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

Кластер-анализ применяется во многих областях, включая компьютерное зрение, маркетинг, биоинформатику и медицину[1].

1.2 Описание возможностей

Программа INDACT предоставляет пользователю возможности просмотра таблиц данных, нормализации данных, кластер-анализа и визуализации результатов. Кроме того, возможности программы включают в себя генерацию искусственных 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 выкачать 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 выделения из таблиц наблюдения множеств (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. Функционалом системы предусмотрено два типа работ:

- кластеризация реальных данных
- проведение численного эксперимента с синтетическими данными

3 Условия применения и 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 Основные принципы работы

4.1 Этапы работы с программой

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 исходных данных
3. Нормализация
4. Отбор признаков
5. Выполнение кластеризации
6. Просмотр результатов и 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 таблицу параметров смартфонов, продаваемых в магазине Ozon (<http://www.ozon.ru/>) в IV квартале 2017 года. Каждому смартфону соответствует 7 параметров: `name`, `price`, `diag`, `cpu`, `ram`, `stype`, `vendor`; соответственно название смартфона, цена в рублях, диагональ экрана в дюймах, частота процессора в ГГц, объем ОЗУ в Мб, тип матрицы, вендор.

Пример файла исходных данных, удовлетворяющий требованиям, описанным в разделе [4.2 Требования к файлу исходных данных](#), приведён ниже. Показаны только несколько первых строк, полный файл содержит 581 модель смартфона. 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

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 разброс, предусмотрены следующие способы вычисления:

- полуразмах
- стандартное отклонение
- абсолютное отклонение

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

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 интегральной таблицы или в виде текстового отчёта.

4.6 Общие сведения о пользовательском интерфейсе

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.

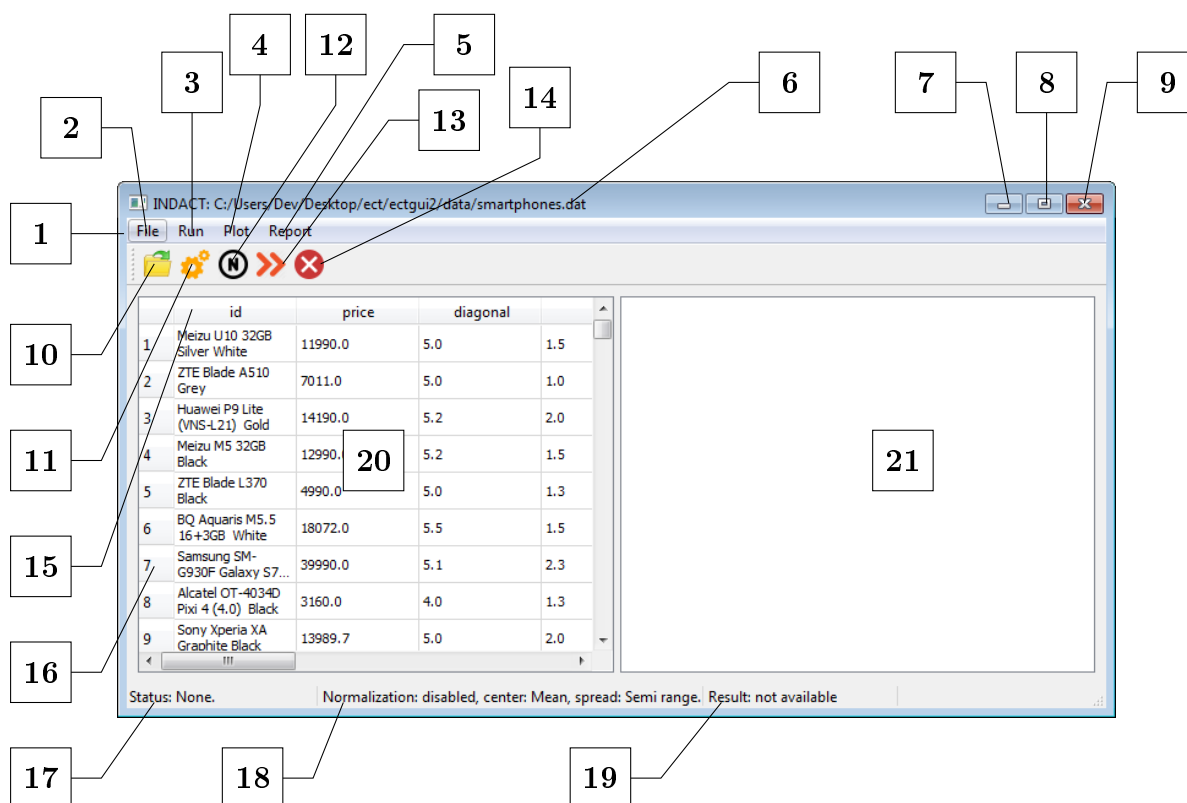


Рисунок 1 – Основные элементы пользовательского интерфейса

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
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`.

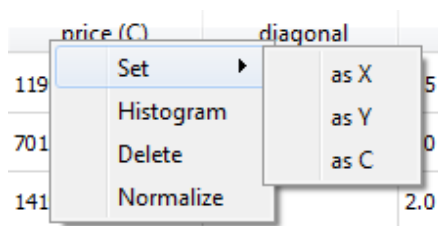


Рисунок 2 – Контекстное меню признака `price`

Контекстное меню содержит следующие пункты:

1. Set — устанавливает особые свойства для признака (см. 5.6.2, 5.4.3)
 - 1.1. as X — выставить метку X для признака (см. 5.6.2)
 - 1.2. as Y — выставить метку Y для признака
 - 1.3. as C — выставить метку C для признака
 - 1.4. as Index (TODO!) — установить признак как индекс (см. 5.4.3)
2. Histogram — строит гистограмму по выбранному признаку (см. 5.6.1)
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 для проведения нормализации (см. 4.4, 5.3).

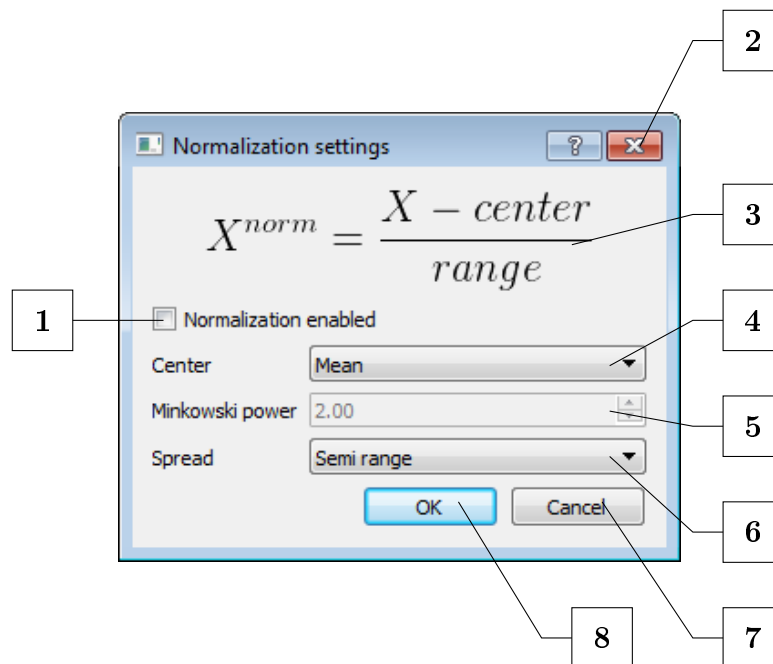


Рисунок 3 – Диалог установки параметров нормализации

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 5.6.1), SVD диаграммы (5.6.3) или поля рассеяния (5.6.2).

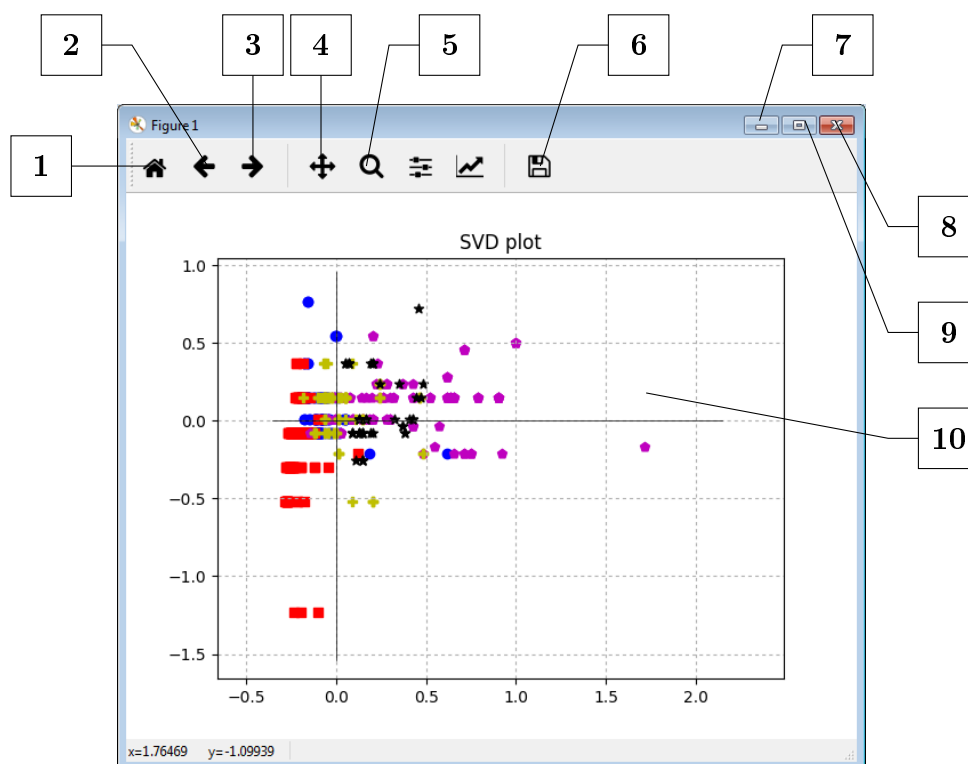


Рисунок 4 – Окно графического вывода

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. Заккрыть окно

10. Изображение

4.6.5 Окно генерации данных

Окно генерации применяется при работе с синтетическими данными (см. раздел 5.7). Это окно необходимо для ввода информации о значениях характеристик генерируемых данных.

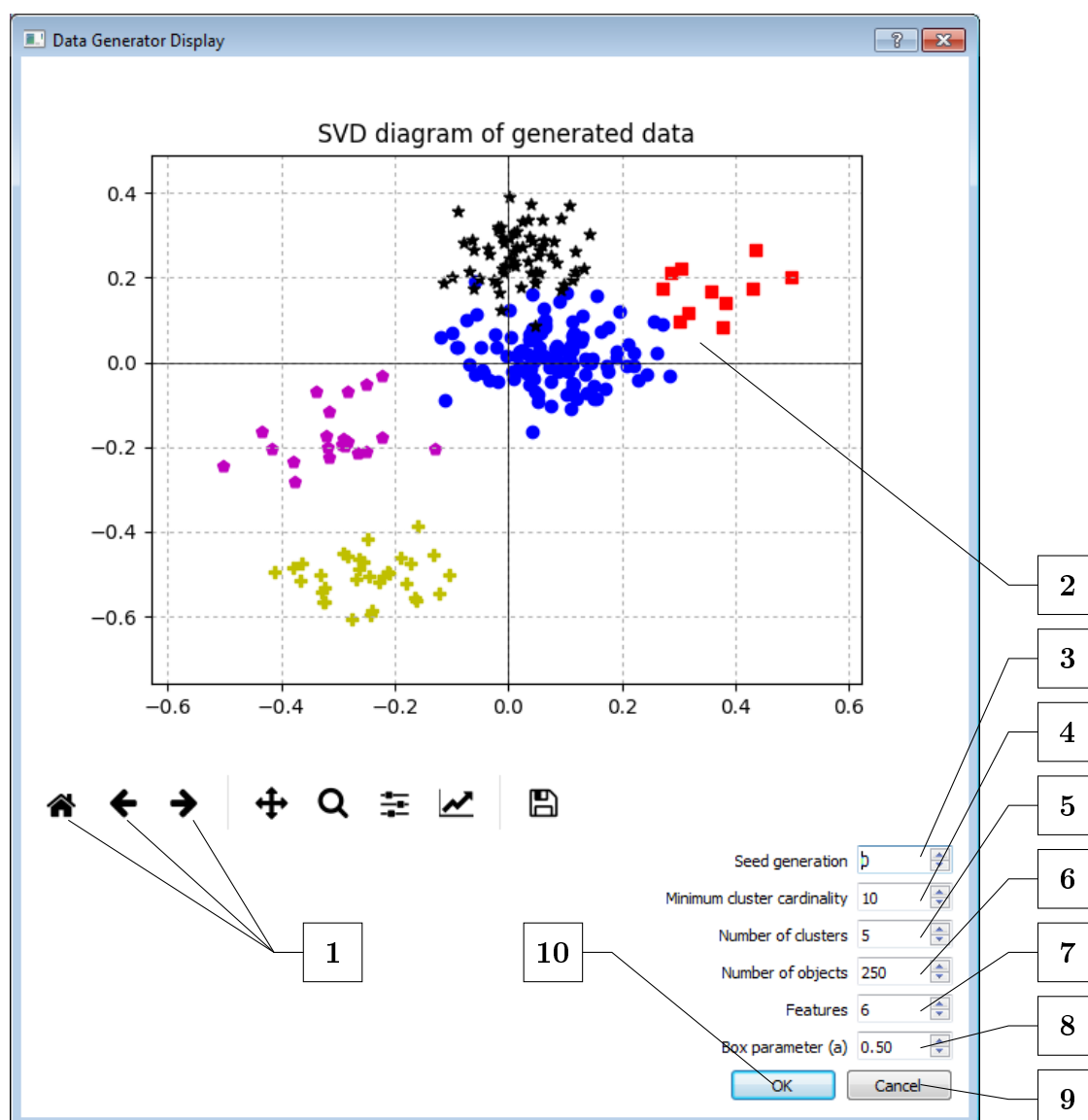


Рисунок 5 – Окно генерации данных

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

Для запуска кластеризации от пользователя требуется выставить ряд параметров в соответствии со значениями которых после будет автоматически выбран подходящий алгоритм кластеризации. На рисунке 6 показано окно для ввода этих параметров.

1. Группа опций, отвечающая за число кластеров
2. Переключатель определения числа кластеров по критерию A-Ward (подробнее см. [3])
3. Переключатель поиска числа кластеров в процессе кластеризации по принципу минимума функции плотности [4]
4. Переключатель явного задания числа кластеров
5. Поле ввода для явного задания числа кластеров
6. Группа опций для установки степени Минковского (см. [1])
7. Переключатель, задающий степень Минковского равной 2
8. Переключатель явного ввода степени Минковского с клавиатуры
9. Поле ввода для степени Минковского
10. Автоматически вычисляемая степень Минковского (в будущих версиях)

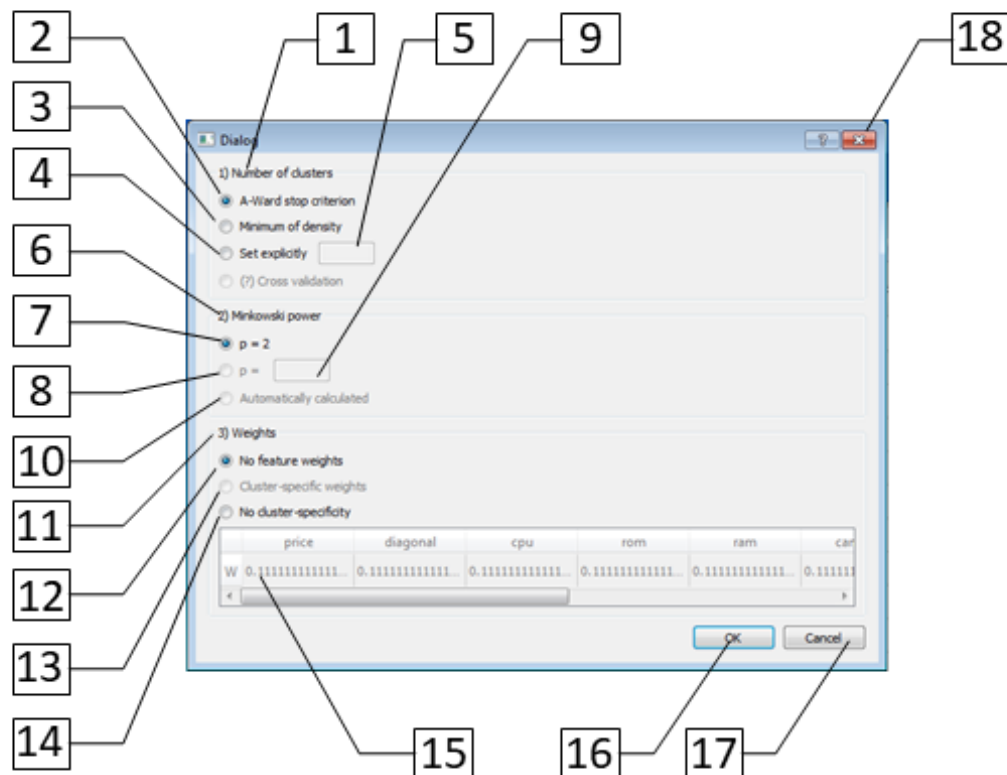


Рисунок 6 – Окно запуска кластеризации

11. Группа опций, определяющих веса признаков
12. Переключатель отключающий веса признаков
13. Переключатель включающий индивидуальные веса для каждого признака (см. [1], $A - Ward_{p\beta}$)
14. Переключатель включающий одинаковые веса в пределах всех данных, но задаваемых для каждого признака индивидуально
15. Таблица для ввода весов признаков
16. Кнопка подтверждения ввода
17. Кнопка отмены

18. Кнопка закрытия окна

4.6.7 Окно таблицы результатов

Окно таблицы результатов служит для интегрального представления полученной кластерной структуры. Это окно может быть вызвано после выполнения шага кластеризации (см. 5.9). Значение таблицы 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.

	price	diag	cpu	ram	Entities
0	7990.000	2.400	1.000	426.667	3
1	41730.067	5.447	2.347	4073.495	91
2	8344.693	5.058	1.291	1565.803	335
3	22423.186	5.300	1.913	2950.737	152
Mean	20121.987	4.551	1.638	2254.175	581

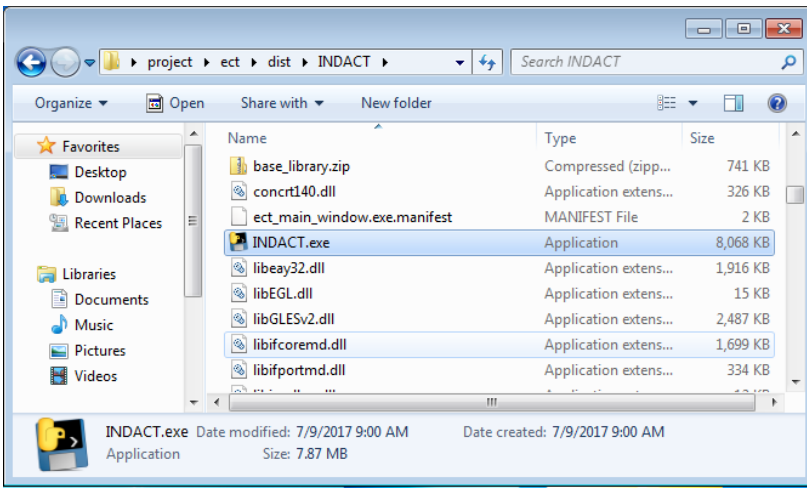
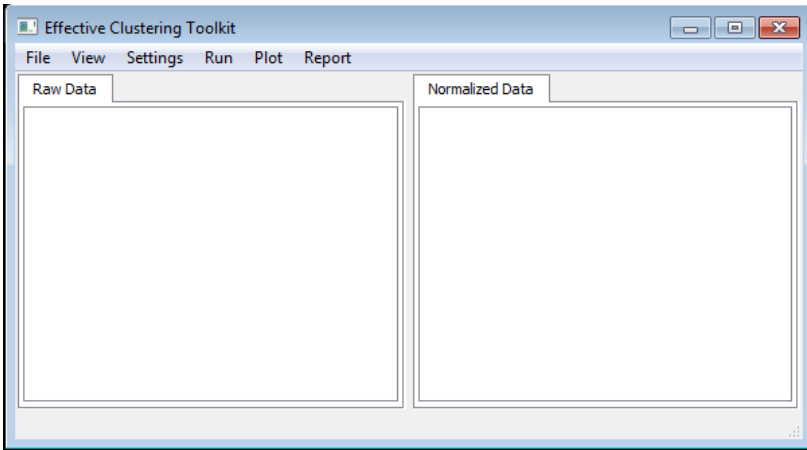
Рисунок 7 – Окно таблицы результатов

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

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>1. <i>Запустить бинарный файл программы</i></p> <p>Дважды нажать левой кнопкой мыши (ЛКМ) на значке <code>INDACT.exe</code></p>	
<p>2. <i>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.</p>	

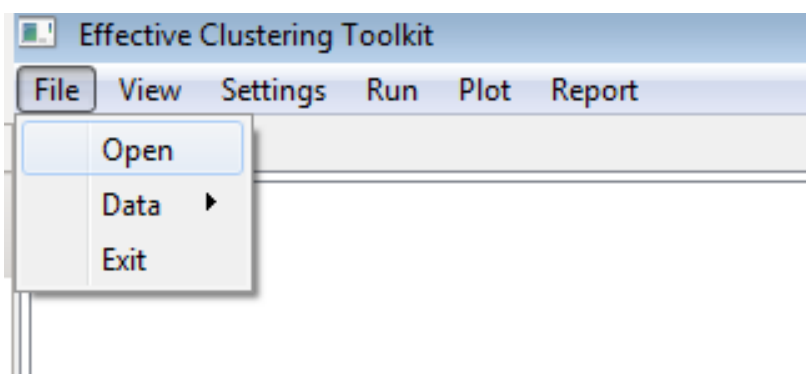
5.2 Загрузка исходных данных

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
-------------------	-----------

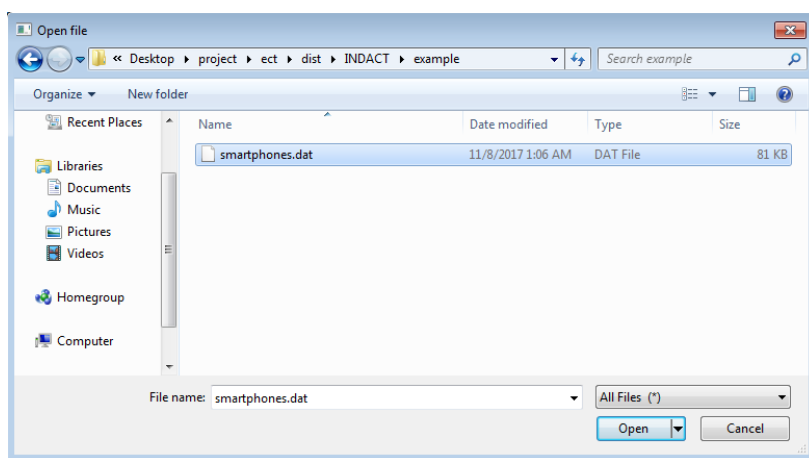
1. Открыть диалог загрузки файла

Для открытия диалога загрузки файла необходимо последовательно нажать в главном меню пункты **File** ⇒ **Open**.



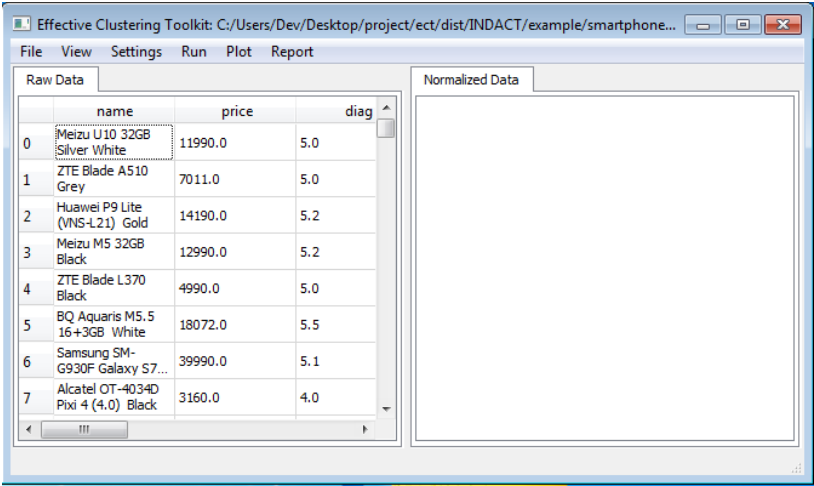
2. Выбрать текстовый файл с данными

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

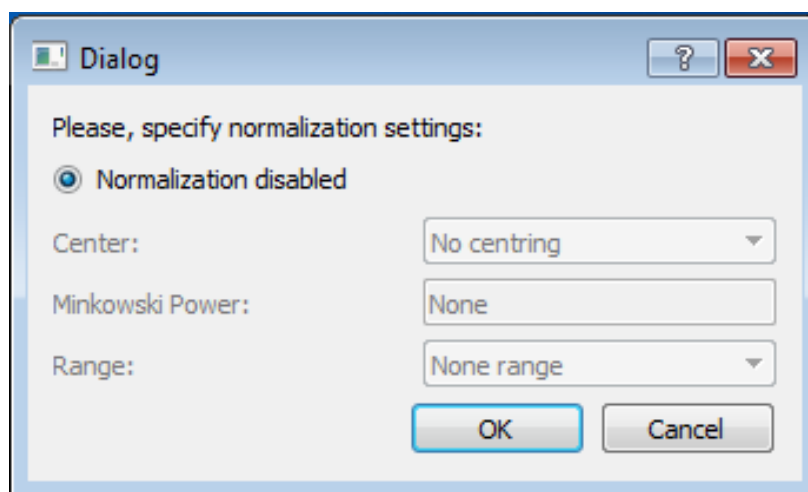
5.3.1 Установка параметров нормализации

Назначение и параметры нормализации описаны в разделе “4.4 Нормализация”.

Действие/Описание	Интерфейс
<p>1. Открыть диалог нормализации</p> <p>Диалог нормализации можно открыть выбрав в главном меню пункты Settings ⇒ Normalization.</p>	The screenshot shows the 'Effective Clustering Toolkit' application window with the 'Settings' menu open. The 'Normalization ...' option is highlighted. The 'Raw Data' tab is still active, showing the same table of smartphone data as in the previous image.

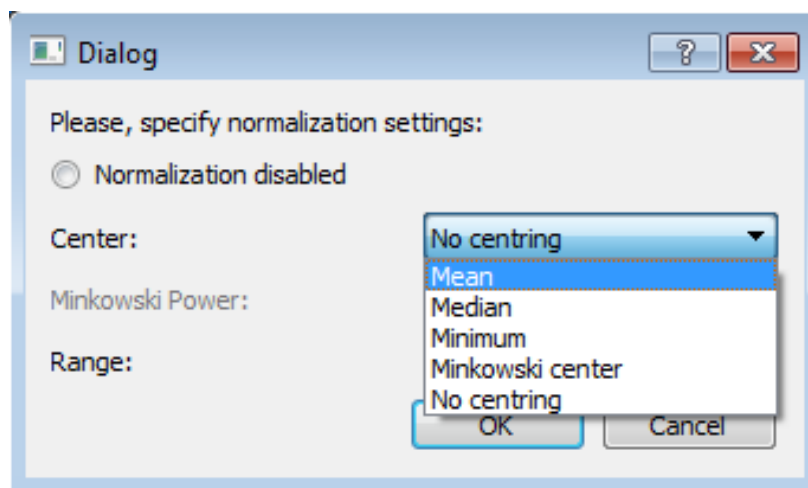
2. Включить нормализацию

Чтобы включить нормализацию требуется снять отметку с пункта “Normalization disabled”. Выполнение этого действия снимет блокировку с полей параметров нормализации.



3. Выставить параметры

Для нормализации данных необходимо задать центр и диапазон нормализации. Эти параметры выбираются из выпадающих списков. Для завершения настройки нормализации нажать кнопку **OK**. Пример установки параметров приведён в разделе 7.1.1.



5.3.2 Нормализация одного признака

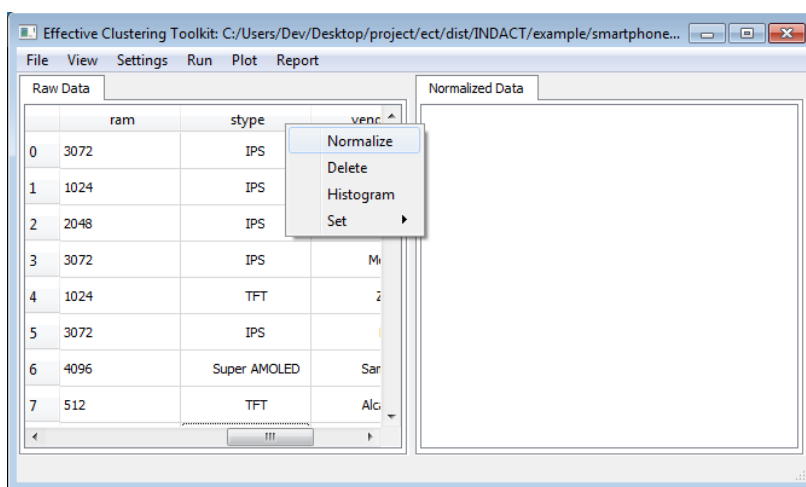
После настройки параметров нормализации необходимо выбрать какие признаки требуется нормализовать. Только выбранные признаки будут участвовать в кластеризации. В программе предусмотрено три возможности для выбора признаков: выбор по одному, выбор всех сразу и удаление по одному.

Действие/Описание

Интерфейс

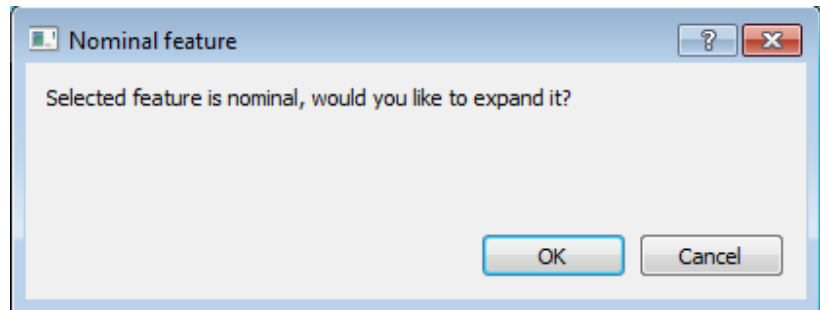
1. Выбрать признак для нормализации

Для выбора одного признака необходимо найти столбец признака во вкладке “Raw Data” 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 произойдёт 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 “Raw Data” во вкладку “Normalized Data” и к нему будут применены выбранные настройки нормализации. Дополнительно во вкладке “Normalized Data” будет отображён столбец “Cluster#”, который будет оставаться заполненным символами “?” до тех пор, пока не будет выполнен шаг кластеризации (см. верхний рисунок, нормализация признака `price`). Сказанное выше справедливо и для номинального признака (например `stype`), но стоит иметь ввиду что соответствующие бинарные признаки будут названы `stype[значение признака]` (как показано на нижнем рисунке)

	name	price	diag
0	Meizu U10 32GB Silver White	11990.0	5.0
1	ZTE Blade A510 Grey	7011.0	5.0
2	Huawei P9 Lite (VNS-L21) Gold	14190.0	5.2
3	Meizu M5 32GB Black	12990.0	5.2
4	ZTE Blade L370 Black	4990.0	5.0
5	BQ Aquaris M5.5 16+3GB White	18072.0	5.5
6	Samsung SM-G930F Galaxy S7...	39990.0	5.1
7	Alcatel OT-4034D Pixi 4 (4.0) Black	3160.0	4.0
8	Sony Xperia XA Graphite Black	13989.7	5.0
9	ZTE Blade L5 Plus Black	6790.0	5.0
10	Meizu Pro 6 64GB		

	price	Cluster#
0	-0.10132177759...	?
1	-0.19558442183...	?
2	-0.05967128205...	?
3	-0.08238973416...	?
4	-0.23384608160...	?
5	0.013822910545...	?
6	0.428775438446...	?
7	-0.26849172108...	?
8	-0.06346337034...	?
9	-0.19976840343...	?
10	0.163726830425...	?

	ram	stype	vendor
0	3072	IPS	Meizu
1	1024	IPS	ZTE
2	2048	IPS	Huawei
3	3072	IPS	Meizu
4	1024	TFT	ZTE
5	3072	IPS	BQ
6	4096	Super AMOLED	Samsung
7	512	TFT	Alcatel
8	2048	IPS	Sony
9	1024	IPS	ZTE

	stypel[IPS]	stypel[TFT]	stypel[S]
0	0.788296041308...	-0.1686746987951807	-0.223752:
1	0.788296041308...	-0.1686746987951807	-0.223752:
2	0.788296041308...	-0.1686746987951807	-0.223752:
3	0.788296041308...	-0.1686746987951807	-0.223752:
4	-1.21170395869...	1.8313253012048194	-0.223752:
5	0.788296041308...	-0.1686746987951807	-0.223752:
6	-1.21170395869...	-0.1686746987951807	1.7762478
7	-1.21170395869...	1.8313253012048194	-0.223752:
8	0.788296041308...	-0.1686746987951807	-0.223752:
9	0.788296041308...	-0.1686746987951807	-0.223752:

5.3.3 Нормализация всех признаков сразу

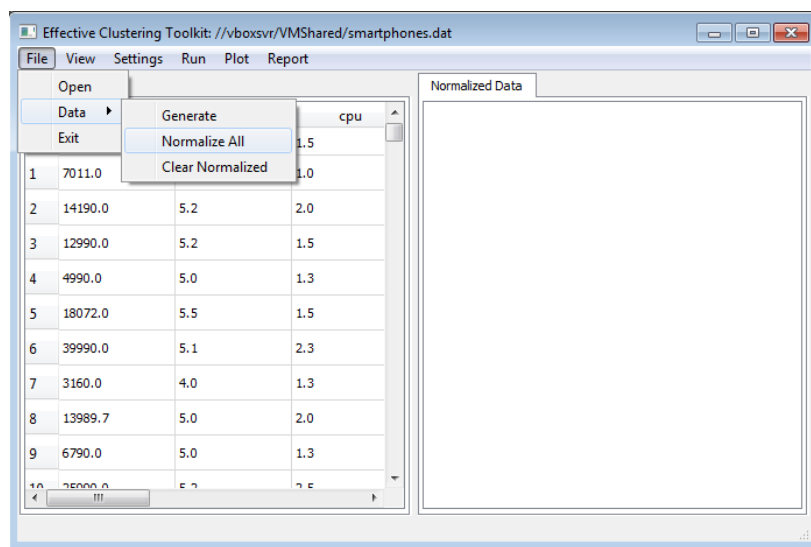
Если признаков много и нормализовать их по одному долго, то можно воспользоваться функцией нормализации всех признаков сразу.

Действие/Описание

Интерфейс

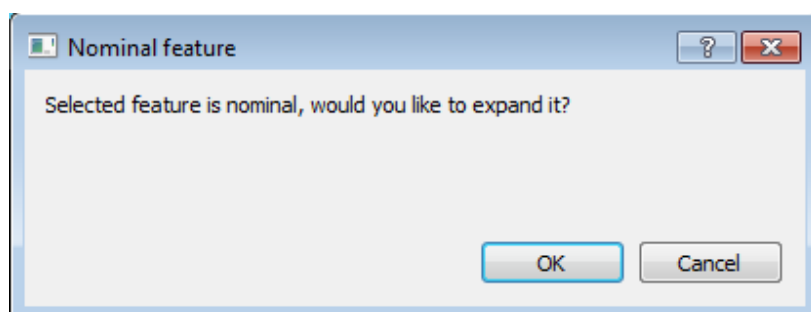
1. Запустить нормализацию всех признаков

Для запуска нормализации всех признаков сразу, требуется в главном меню выбрать **File** ⇒ **Data** ⇒ **Normalize All**.



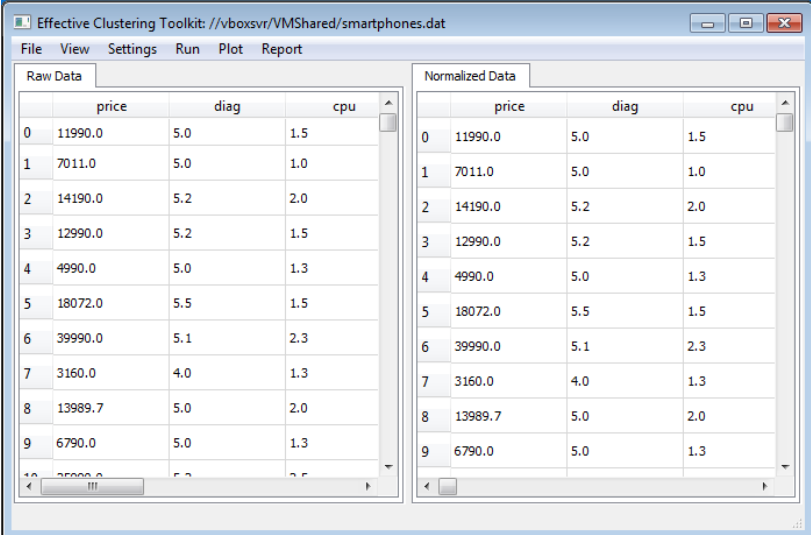
2. (При необходимости) Подтвердить нормализацию категориального признака

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 во вкладке “Normalized Data”



The screenshot shows the 'Effective Clustering Toolkit' application window. It has a menu bar with 'File', 'View', 'Settings', 'Run', 'Plot', and 'Report'. Below the menu bar are two tabs: 'Raw Data' and 'Normalized Data'. Both tabs display a table with 11 rows and 4 columns: an index column, 'price', 'diag', and 'cpu'. The data in both tables is identical.

	price	diag	cpu
0	11990.0	5.0	1.5
1	7011.0	5.0	1.0
2	14190.0	5.2	2.0
3	12990.0	5.2	1.5
4	4990.0	5.0	1.3
5	18072.0	5.5	1.5
6	39990.0	5.1	2.3
7	3160.0	4.0	1.3
8	13989.7	5.0	2.0
9	6790.0	5.0	1.3
10	35000.0	5.0	2.5

5.4 Отбор признаков

5.4.1 Удаление одного признака

Как было отмечено выше, программа позволяет удалять отдельные признаки из как вкладки “Normalized Data” так и “Raw Data”. Эта функция может быть применена для исключения из рассмотрения определённых признаков.

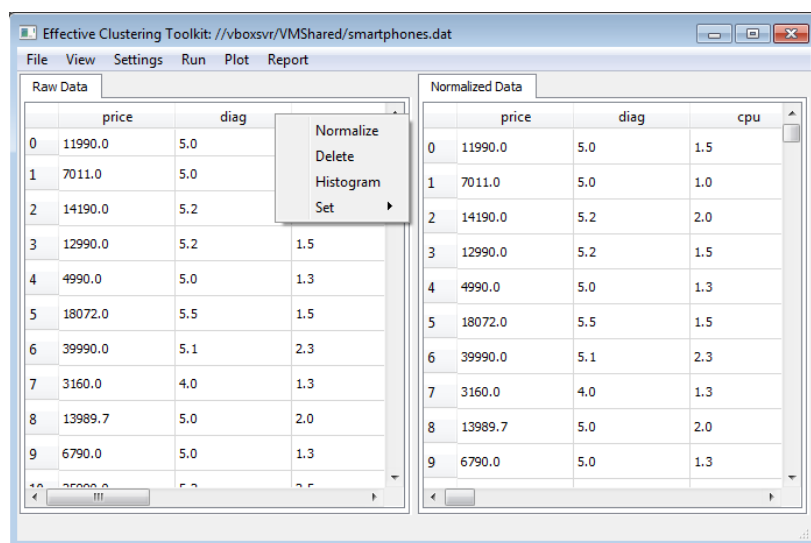
Действие/Описание

Интерфейс

1. Выбрать признак для удаления

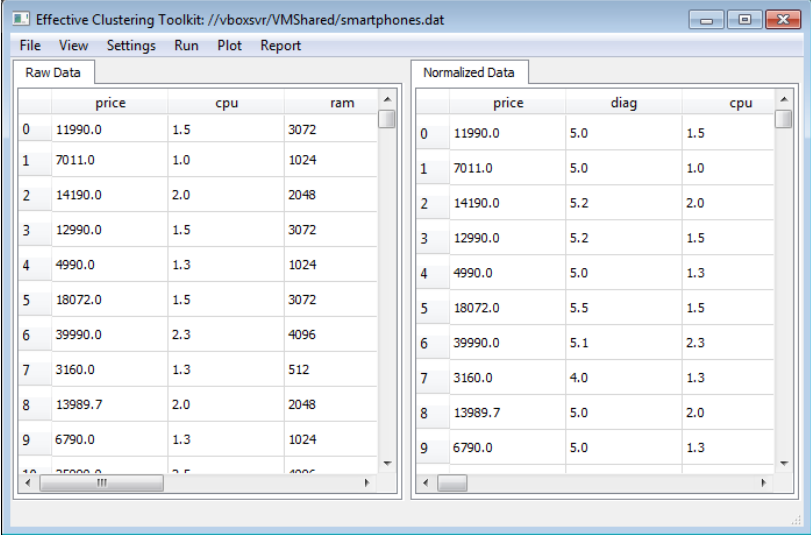
Для удаления одного признака необходимо найти столбец признака в нужной вкладке и нажать на нём ПКМ. В контекстном меню выбрать пункт **Delete**.

Рассмотрим удаление на примере признака **diag**. Контекстное меню, открытое после нажатия на заголовке **diag** показано на рисунке справа.



2. Посмотреть результат

В результате выполнения операции выбранный признак будет удалён только из вкладки “Raw Data”, но останется во второй вкладке.



The screenshot shows the 'Effective Clustering Toolkit' window with the file path '//vboxsvr/VMShared/smartphones.dat'. The interface has two tabs: 'Raw Data' and 'Normalized Data'. The 'Raw Data' tab shows a table with columns 'price', 'cpu', and 'ram'. The 'Normalized Data' tab shows a table with columns 'price', 'diag', and 'cpu'. The 'cpu' column is present in both, while 'ram' is only in 'Raw Data' and 'diag' is only in 'Normalized Data'.

	price	cpu	ram
0	11990.0	1.5	3072
1	7011.0	1.0	1024
2	14190.0	2.0	2048
3	12990.0	1.5	3072
4	4990.0	1.3	1024
5	18072.0	1.5	3072
6	39990.0	2.3	4096
7	3160.0	1.3	512
8	13989.7	2.0	2048
9	6790.0	1.3	1024

	price	diag	cpu
0	11990.0	5.0	1.5
1	7011.0	5.0	1.0
2	14190.0	5.2	2.0
3	12990.0	5.2	1.5
4	4990.0	5.0	1.3
5	18072.0	5.5	1.5
6	39990.0	5.1	2.3
7	3160.0	4.0	1.3
8	13989.7	5.0	2.0
9	6790.0	5.0	1.3

5.4.2 Удаление всех признаков сразу

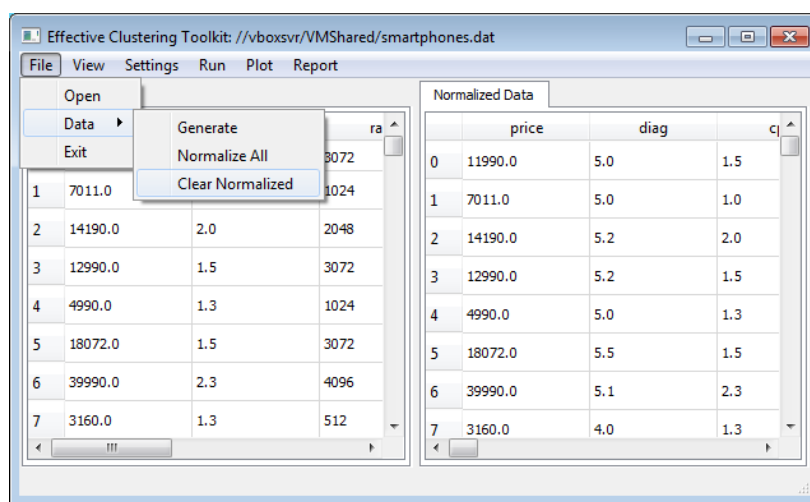
Если требуется полностью очистить вкладку “Normalized Data”, то следует воспользоваться функцией, описанной ниже. Функция очистки вкладки нормализованных данных применяется для того чтобы сбросить выбор всех признаков.

Действие/Описание

Интерфейс

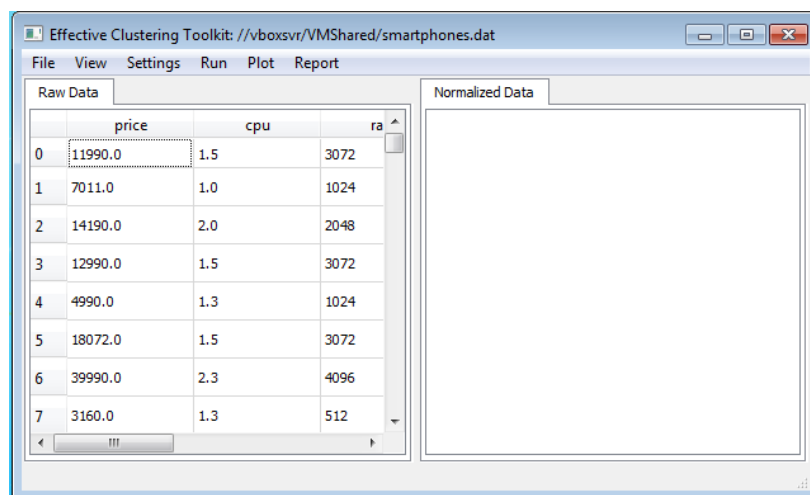
1. Запустить удаление

Функция удаления всех признаков сразу вызывается из главного меню программы: **File** ⇒ **Data** ⇒ **Clear Normalized**



2. Посмотреть результат

В результате выполнения операции вкладка “Normalized Data” будет очищена полностью.



5.4.3 Установка признака как индекс

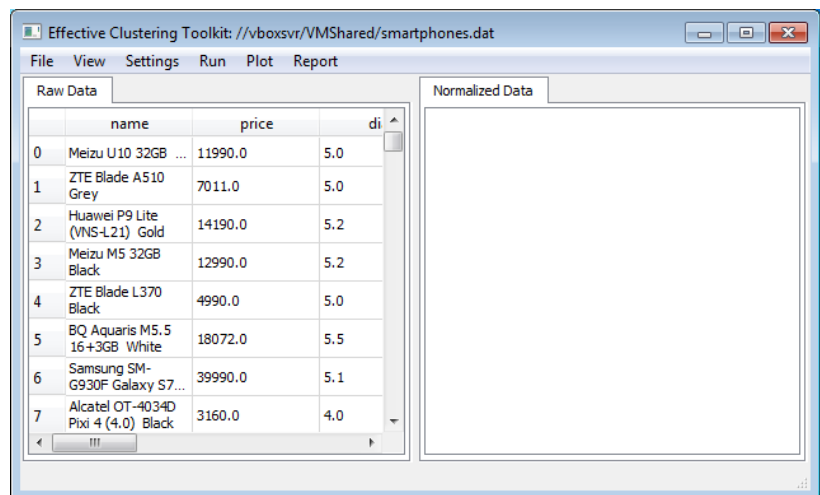
Допустим, исходный файл содержит признак, который не требуется использовать для анализа, но который был бы полезен как уникальный индекс, например, это может быть название модели телефона. В таком случае в программе предусмотрена функция задания признака в качестве индекса.

Действие/Описание

Интерфейс

1. Загрузить файл

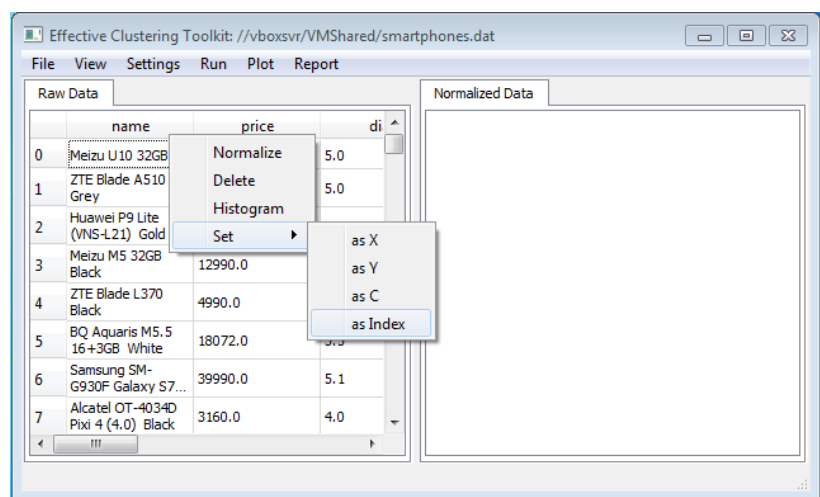
Загрузка файла описана в разделе “5.2 Загрузка исходных данных”.



2. Выбрать признак

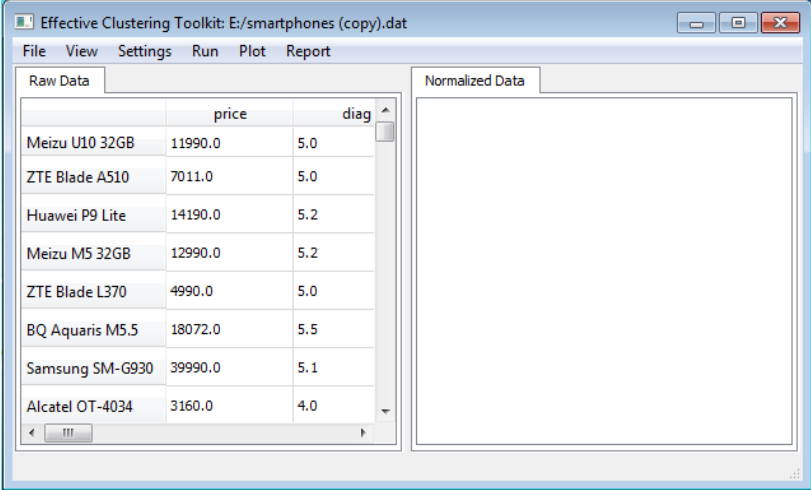
Для выбора признака необходимо найти его столбец в одной из вкладок и нажать ПКМ на нём. В контекстном меню выбрать пункт **Set** ⇒ **As Index**

Set ⇒ **As Index**



3. Посмотреть результат

Результат выполнения предыдущего действия состоит в отображении признака в колонке индекса как показано на рисунке справа.



The screenshot shows the 'Effective Clustering Toolkit' window with the file 'E:/smartphones (copy).dat'. The 'Raw Data' tab is active, displaying a table with three columns: an unnamed column for smartphone models, 'price', and 'diag'. The 'Normalized Data' tab is empty. The table contains the following data:

	price	diag
Meizu U10 32GB	11990.0	5.0
ZTE Blade A510	7011.0	5.0
Huawei P9 Lite	14190.0	5.2
Meizu M5 32GB	12990.0	5.2
ZTE Blade L370	4990.0	5.0
BQ Aquaris M5.5	18072.0	5.5
Samsung SM-G930	39990.0	5.1
Alcatel OT-4034	3160.0	4.0

5.5 Настройка способа отображения вкладок

Программа имеет два способа отображения данных: с помощью вкладок и с помощью панелей (по умолчанию). Для переключения этих способов служит пункт меню

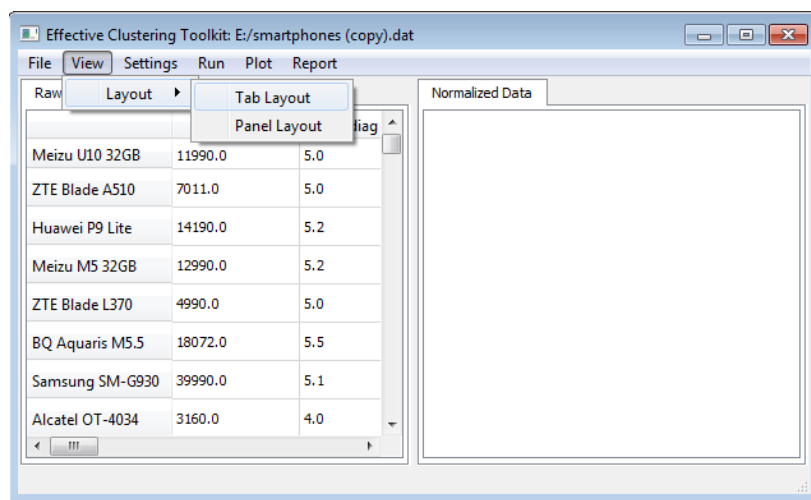
View

Действие/Описание

Интерфейс

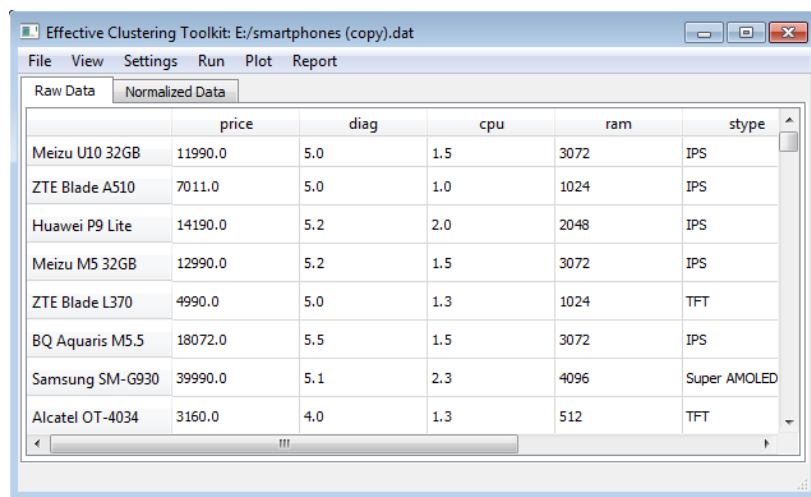
1. Переключить способ отображения

Для переключения способа отображения следует выбрать из главного меню программы: **View** ⇒ **Layout** ⇒ **Tab Layout** или **Panel Layout**



2. Посмотреть результат

В результате выполнения операции панели “Raw Data” и “Normalized Data” будут отображены одна за другой.



5.6 Визуализация

5.6.1 Построение гистограммы по признаку

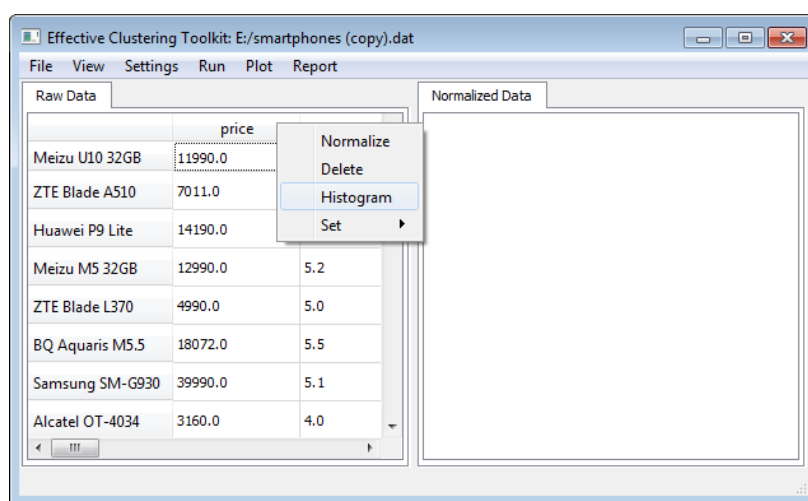
В качестве первичного инструмента анализа программа предлагает возможность построения гистограммы по выбранному признаку.

Действие/Описание

Интерфейс

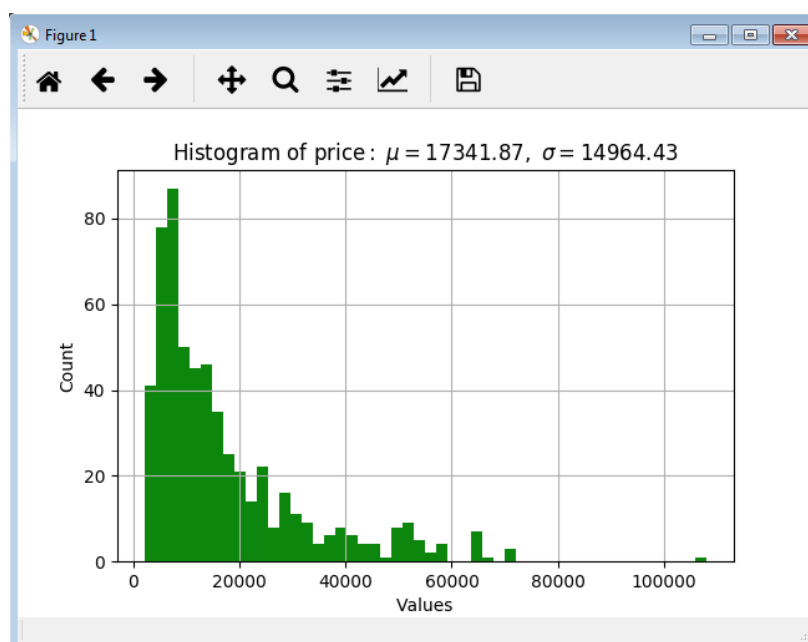
1. Выбрать признак

Для выбора признака необходимо найти его столбец в одной из вкладок и нажать ПКМ на нём. В контекстном меню выбрать пункт **Histogram**. На примере показано построение гистограммы по признаку `price`.



2. Посмотреть результат

После выбора признака будет построена гистограмма в отдельном окне.



5.6.2 Построение поля рассеяния (scatter plot)

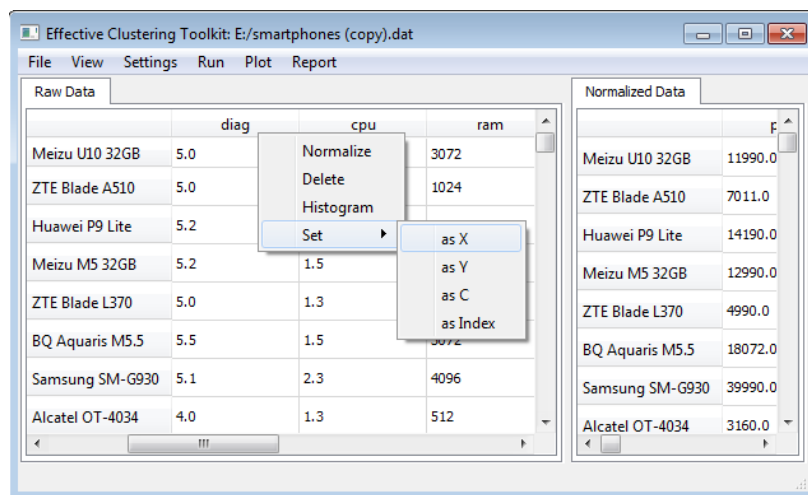
Для первичного анализа структуры данных или результатов кластеризации в программе предусмотрена функция построения поля рассеяния по меткам на выбранных признаках. Метка — вспомогательный символ, присваиваемый пользователем для определённого признака. Предусмотрено 3 вида меток: “X”, “Y”, “C”. Первый 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

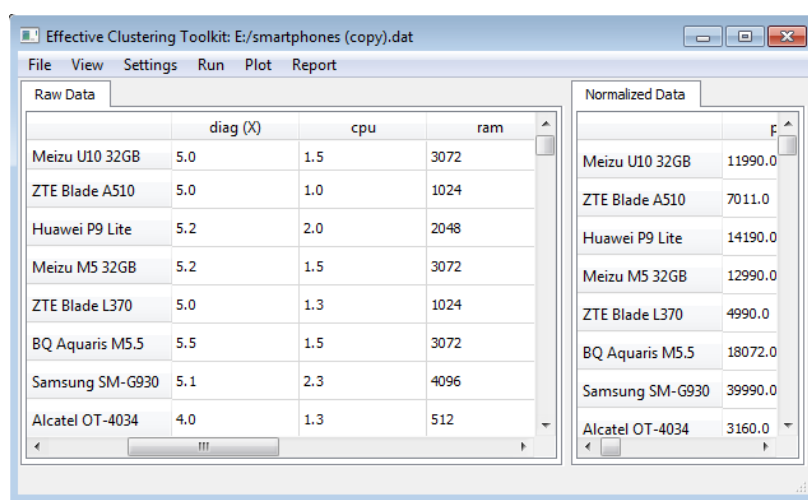
1. Выбрать признак по оси X

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)”

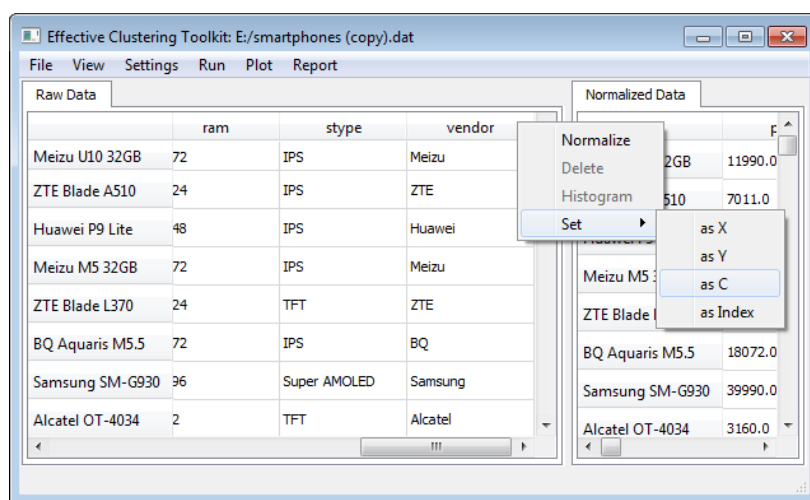


Аналогично пунктам 1,2.

3. Выбрать признак по оси Y

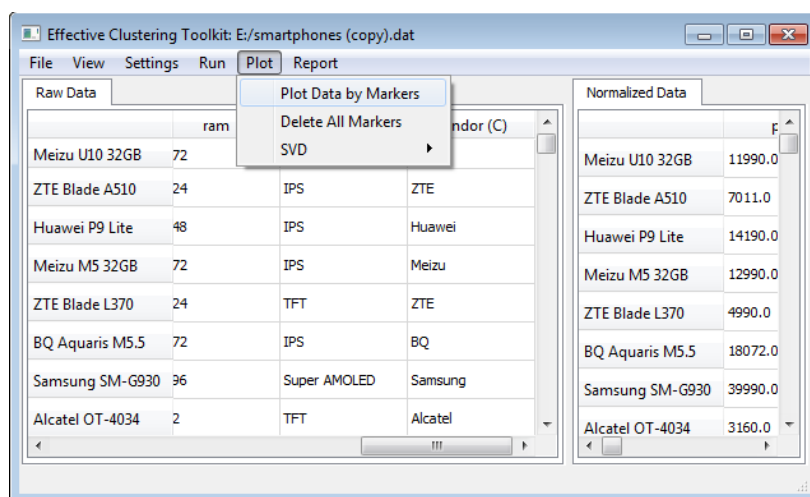
4. Выбрать признак отвечающий за цвет точек

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



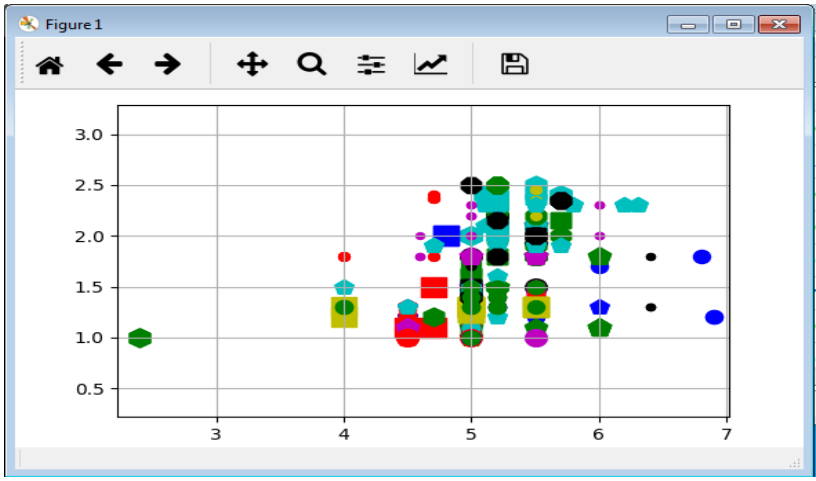
5. Построить scatter plot

В главном меню выбрать **Plot** ⇒ **Plot Data by Markers**



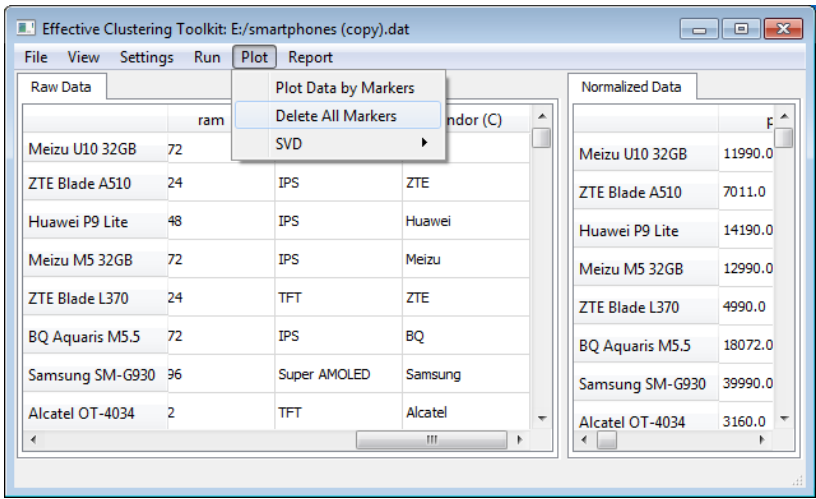
6. Посмотреть
результат

В новом окне откроется по-
строенная диаграмма



7. Удалить метки
(опционально)

В главном меню
выбрать **Plot** ⇒
Delete All Markers.
При этом отметки “(X)”,
“(Y)” и “(C)” будут удале-
ны.



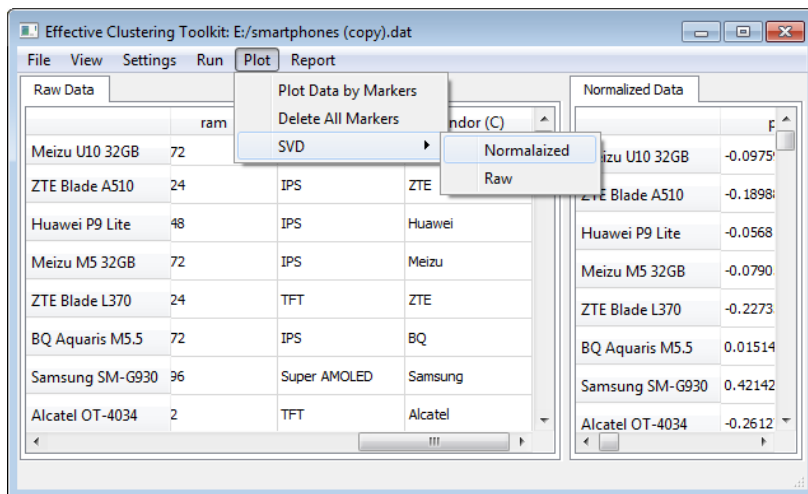
5.6.3 Построение SVD диаграммы

Для интегральной оценки структуры данных предусмотрена функция построения SVD диаграммы. Имеется возможность построения SVD диаграммы по нормализован-
ным и не нормализованным данным.

Действие/Описание	Интерфейс
-------------------	-----------

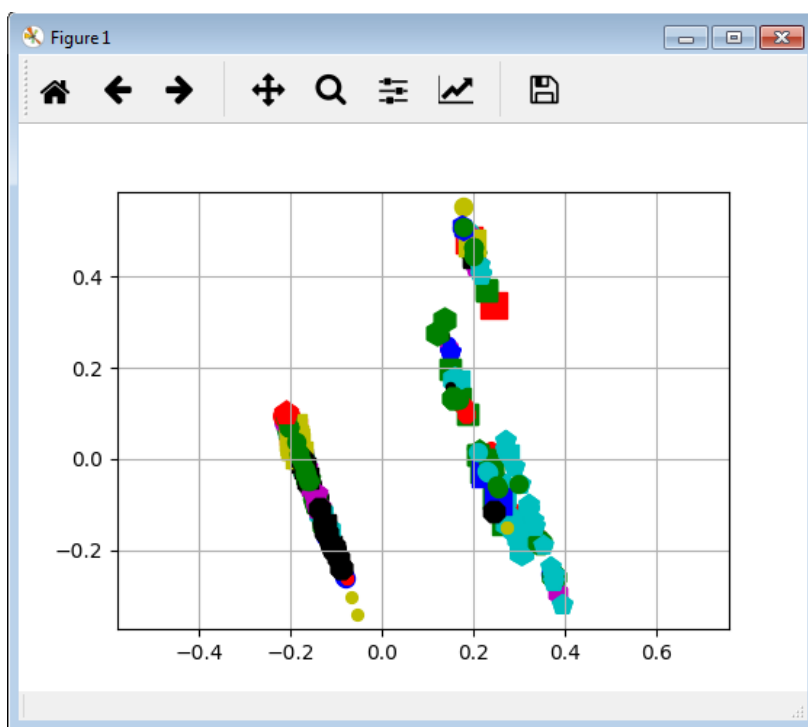
1. Построить SVD диаграмму

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.



5.7 Генерация синтетических данных

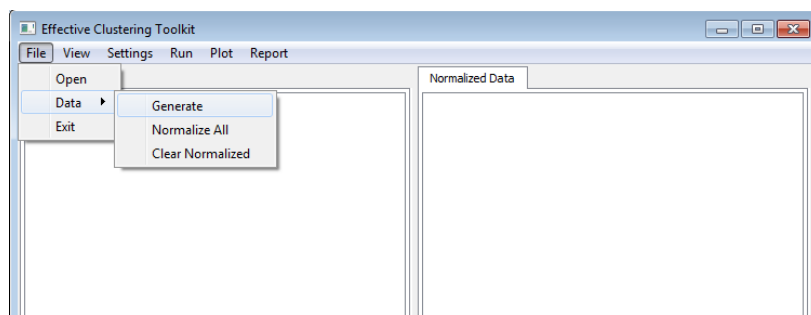
Для генерирования искусственных данных необходимо вызвать диалог настройки параметров, указать все необходимые величины и сохранить сгенерированные данные в файл.

Действие/Описание

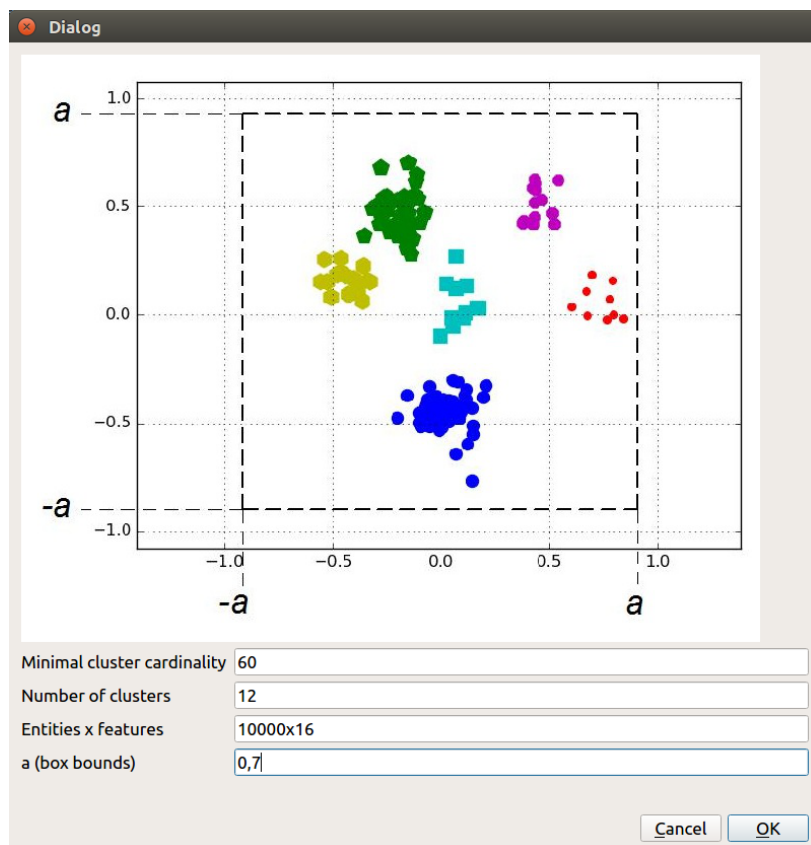
Интерфейс

1. Открыть диалог генерации данных

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

*2. Указать параметры данных*

В открывшемся диалоге необходимо указать параметры данных, по которым будет производиться генерация. Подробнее о параметрах генерации см. [4]. В верхней части диалога отображается статическая информирующая диаграмма. Когда все параметры будут введены, нажать кнопку **OK** и в стандартном диалоге сохранения указать файл, в который требуется записать результат.



-
3. *Сохранить данные в файл* Сохранить данные в файл в стандартном файловом диалоге.
-

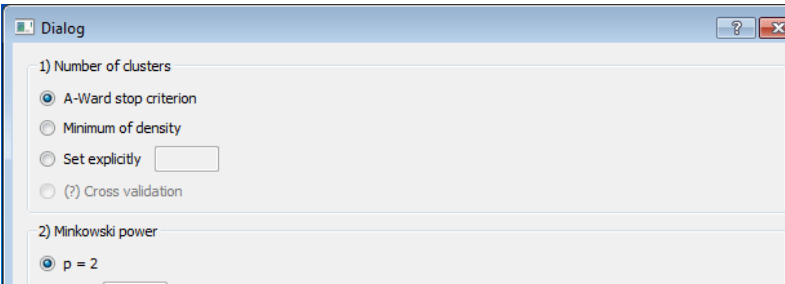
5.8 Запуск кластеризации

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

Действие/Описание
Интерфейс

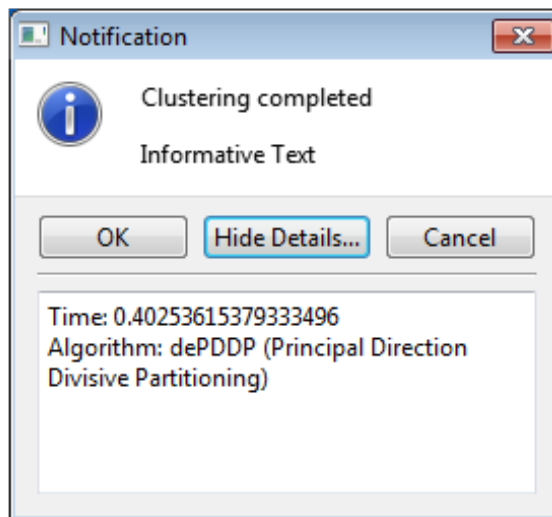
1. Открыть диалог выбора параметров

Пред запуском кластеризации потребуется задать общие параметры процедуры, по которым программа выберет конкретный алгоритм. Для этого необходимо из главного меню выбрать: **Run** ⇒ **Clustering**. Подробнее про параметры см. [1, 4] и раздел 6 После установки выбранных параметров для подтверждения нажать кнопку **OK**



2. Дождаться результатов кластеризации

Сразу после нажатия кнопки **OK** начнется работа алгоритма кластеризации. Когда алгоритм закончит работу, появится окно, изображённое справа. Если в окне нажать кнопку **Show Details** то появится дополнительная информация о выбранном алгоритме и времени работы. Нажать кнопку **OK**.



3. Проверить заполнение столбца Cluster#

В процессе кластеризации каждому объекту ставится в соответствие номер кластера, которому он принадлежит. Для заданного объекта этот номер можно посмотреть в столбце "Cluster#"

Raw Data			
	price	diag	cpu
0	11990.0	5.0	1.5
1	7011.0	5.0	1.0
2	14190.0	5.2	2.0
3	12990.0	5.2	1.5
4	4990.0	5.0	1.3
5	18072.0	5.5	1.5
6	39990.0	5.1	2.3
7	3160.0	4.0	1.3
8	13989.7	5.0	2.0
9	6790.0	5.0	1.3
10	25990.0	5.2	2.5

Normalized Data			Cluster#
cpu	ram		
0	-0.15674125071...	0.190772305813...	2
1	-0.82340791738...	-0.32535672644...	2
2	0.509925415949...	-0.06729221031...	3
3	-0.15674125071...	0.190772305813...	2
4	-0.42340791738...	-0.32535672644...	2
5	-0.15674125071...	0.190772305813...	2
6	0.909925415949...	0.448836821942...	1
7	-0.42340791738...	-0.45438898450...	2
8	0.509925415949...	-0.06729221031...	3
9	-0.42340791738...	-0.32535672644...	2
10	1.176592082616...	0.448836821942...	1

5.9 Генерация отчёта

Результаты кластеризации удобно анализировать по сгенерированному отчёту.

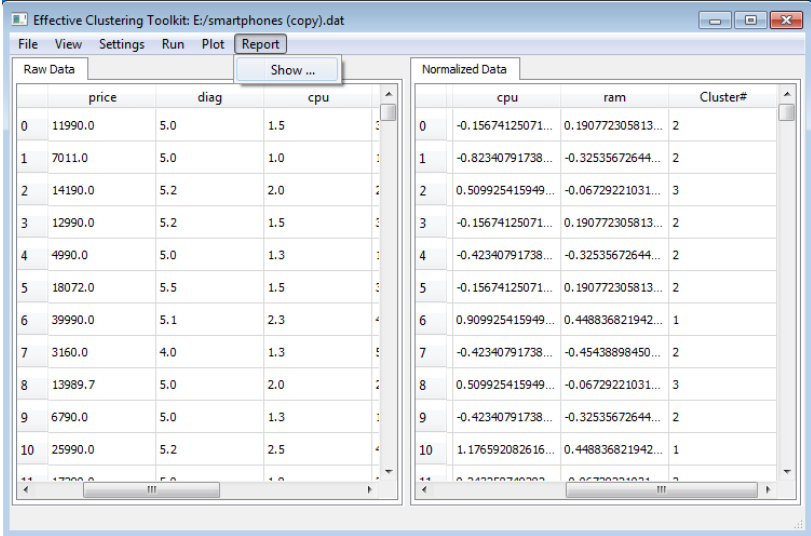
Действие/Описание

Интерфейс

1. Сгенерировать отчёт

Для генерации отчёта в главном меню выбрать

Report ⇒ **Show**.



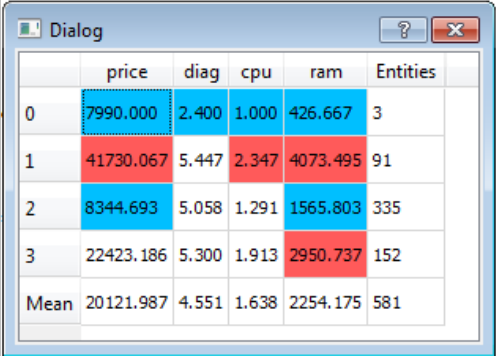
The screenshot shows the 'Effective Clustering Toolkit' window with the 'Report' menu item highlighted. The 'Raw Data' table on the left contains columns for 'price', 'diag', and 'cpu'. The 'Normalized Data' table on the right contains columns for 'cpu', 'ram', and 'Cluster#'. Both tables display 11 rows of data.

Raw Data			
	price	diag	cpu
0	11990.0	5.0	1.5
1	7011.0	5.0	1.0
2	14190.0	5.2	2.0
3	12990.0	5.2	1.5
4	4990.0	5.0	1.3
5	18072.0	5.5	1.5
6	39990.0	5.1	2.3
7	3160.0	4.0	1.3
8	13989.7	5.0	2.0
9	6790.0	5.0	1.3
10	25990.0	5.2	2.5

Normalized Data			
	cpu	ram	Cluster#
0	-0.15674125071...	0.190772305813...	2
1	-0.82340791738...	-0.32535672644...	2
2	0.509925415949...	-0.06729221031...	3
3	-0.15674125071...	0.190772305813...	2
4	-0.42340791738...	-0.32535672644...	2
5	-0.15674125071...	0.190772305813...	2
6	0.909925415949...	0.448836821942...	1
7	-0.42340791738...	-0.45438898450...	2
8	0.509925415949...	-0.06729221031...	3
9	-0.42340791738...	-0.32535672644...	2
10	1.176592082616...	0.448836821942...	1

*2. Посмотреть окно
таблицы результатов*

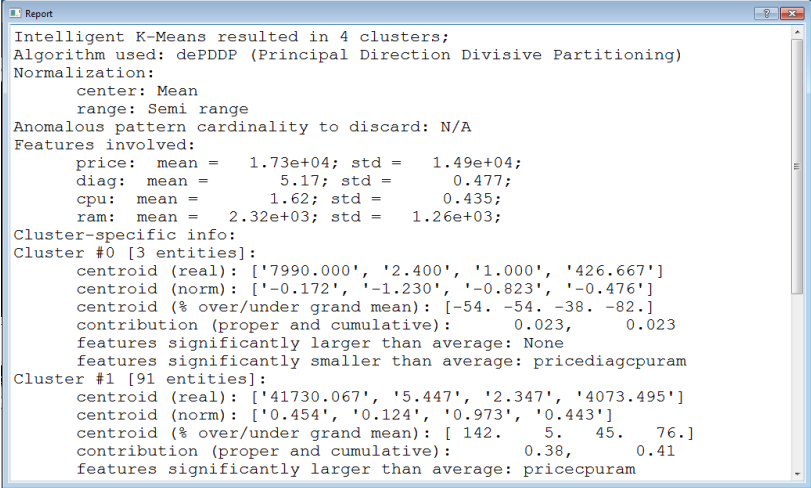
Отчёт состоит из двух окон. Первое — окно с таблицей результатов. В этом окне приведена сводная таблица в которой строки соответствуют кластерам, а столбцы — признакам. В ячейках указаны средние значения признака по кластеру. Красным цветом выделены ячейки, в которых относительная разность значения и средней величины признака по кластеру больше 30%, соответственно синим — меньше 30%. Маржинальная строка содержит средние значения признаков по всем кластерам, а столбец — число объектов в кластере.



	price	diag	cpu	ram	Entities
0	7990.000	2.400	1.000	426.667	3
1	41730.067	5.447	2.347	4073.495	91
2	8344.693	5.058	1.291	1565.803	335
3	22423.186	5.300	1.913	2950.737	152
Mean	20121.987	4.551	1.638	2254.175	581

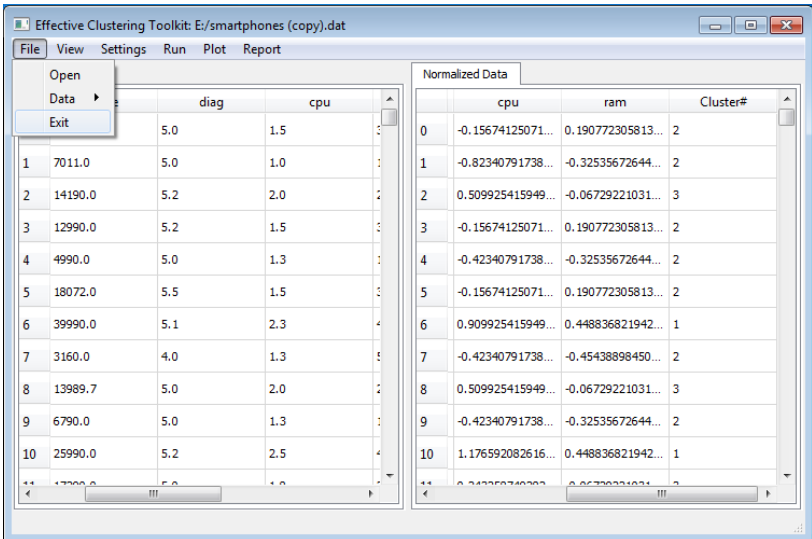
3. Посмотреть окно текстового отчёта

Текстовый отчёт содержит все сведения относительно выбранного алгоритма, метода нормализации, и параметров каждого из кластеров.



```
Report
Intelligent K-Means resulted in 4 clusters;
Algorithm used: dePDDP (Principal Direction Divisive Partitioning)
Normalization:
  center: Mean
  range: Semi range
Anomalous pattern cardinality to discard: N/A
Features involved:
  price: mean = 1.73e+04; std = 1.49e+04;
  diag: mean = 5.17; std = 0.477;
  cpu: mean = 1.62; std = 0.435;
  ram: mean = 2.32e+03; std = 1.26e+03;
Cluster-specific info:
Cluster #0 [3 entities]:
  centroid (real): ['7990.000', '2.400', '1.000', '426.667']
  centroid (norm): ['-0.172', '-1.230', '-0.823', '-0.476']
  centroid (% over/under grand mean): [-54. -54. -38. -82.]
  contribution (proper and cumulative): 0.023, 0.023
  features significantly larger than average: None
  features significantly smaller than average: pricediagcpuram
Cluster #1 [91 entities]:
  centroid (real): ['41730.067', '5.447', '2.347', '4073.495']
  centroid (norm): ['0.454', '0.124', '0.973', '0.443']
  centroid (% over/under grand mean): [ 142. 5. 45. 76.]
  contribution (proper and cumulative): 0.38, 0.41
  features significantly larger than average: pricecpuram
```

5.10 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.</p>	 <p>The screenshot shows the 'Effective Clustering Toolkit' window with the title 'E:\smartphones (copy).dat'. The 'File' menu is open, showing options: 'Open', 'Data', and 'Exit'. The 'Exit' option is highlighted. The main window displays two data tables. The left table has columns 'diag' and 'cpu'. The right table, titled 'Normalized Data', has columns 'cpu', 'ram', and 'Cluster#'. Both tables show 10 rows of data.</p> <table><tr><th></th><th>diag</th><th>cpu</th></tr><tr><td>1</td><td>7011.0</td><td>5.0</td></tr><tr><td>2</td><td>14190.0</td><td>5.2</td></tr><tr><td>3</td><td>12990.0</td><td>5.2</td></tr><tr><td>4</td><td>4990.0</td><td>5.0</td></tr><tr><td>5</td><td>18072.0</td><td>5.5</td></tr><tr><td>6</td><td>39990.0</td><td>5.1</td></tr><tr><td>7</td><td>3160.0</td><td>4.0</td></tr><tr><td>8</td><td>13989.7</td><td>5.0</td></tr><tr><td>9</td><td>6790.0</td><td>5.0</td></tr><tr><td>10</td><td>25990.0</td><td>5.2</td></tr></table> <table><tr><th></th><th>cpu</th><th>ram</th><th>Cluster#</th></tr><tr><td>0</td><td>-0.15674125071...</td><td>0.190772305813...</td><td>2</td></tr><tr><td>1</td><td>-0.82340791738...</td><td>-0.32535672644...</td><td>2</td></tr><tr><td>2</td><td>0.509925415949...</td><td>-0.06729221031...</td><td>3</td></tr><tr><td>3</td><td>-0.15674125071...</td><td>0.190772305813...</td><td>2</td></tr><tr><td>4</td><td>-0.42340791738...</td><td>-0.32535672644...</td><td>2</td></tr><tr><td>5</td><td>-0.15674125071...</td><td>0.190772305813...</td><td>2</td></tr><tr><td>6</td><td>0.909925415949...</td><td>0.448836821942...</td><td>1</td></tr><tr><td>7</td><td>-0.42340791738...</td><td>-0.45438898450...</td><td>2</td></tr><tr><td>8</td><td>0.509925415949...</td><td>-0.06729221031...</td><td>3</td></tr><tr><td>9</td><td>-0.42340791738...</td><td>-0.32535672644...</td><td>2</td></tr><tr><td>10</td><td>1.176592082616...</td><td>0.448836821942...</td><td>1</td></tr></table>		diag	cpu	1	7011.0	5.0	2	14190.0	5.2	3	12990.0	5.2	4	4990.0	5.0	5	18072.0	5.5	6	39990.0	5.1	7	3160.0	4.0	8	13989.7	5.0	9	6790.0	5.0	10	25990.0	5.2		cpu	ram	Cluster#	0	-0.15674125071...	0.190772305813...	2	1	-0.82340791738...	-0.32535672644...	2	2	0.509925415949...	-0.06729221031...	3	3	-0.15674125071...	0.190772305813...	2	4	-0.42340791738...	-0.32535672644...	2	5	-0.15674125071...	0.190772305813...	2	6	0.909925415949...	0.448836821942...	1	7	-0.42340791738...	-0.45438898450...	2	8	0.509925415949...	-0.06729221031...	3	9	-0.42340791738...	-0.32535672644...	2	10	1.176592082616...	0.448836821942...	1
	diag	cpu																																																																																
1	7011.0	5.0																																																																																
2	14190.0	5.2																																																																																
3	12990.0	5.2																																																																																
4	4990.0	5.0																																																																																
5	18072.0	5.5																																																																																
6	39990.0	5.1																																																																																
7	3160.0	4.0																																																																																
8	13989.7	5.0																																																																																
9	6790.0	5.0																																																																																
10	25990.0	5.2																																																																																
	cpu	ram	Cluster#																																																																															
0	-0.15674125071...	0.190772305813...	2																																																																															
1	-0.82340791738...	-0.32535672644...	2																																																																															
2	0.509925415949...	-0.06729221031...	3																																																																															
3	-0.15674125071...	0.190772305813...	2																																																																															
4	-0.42340791738...	-0.32535672644...	2																																																																															
5	-0.15674125071...	0.190772305813...	2																																																																															
6	0.909925415949...	0.448836821942...	1																																																																															
7	-0.42340791738...	-0.45438898450...	2																																																																															
8	0.509925415949...	-0.06729221031...	3																																																																															
9	-0.42340791738...	-0.32535672644...	2																																																																															
10	1.176592082616...	0.448836821942...	1																																																																															

6 Алгоритмы кластеризации (краткое описание)

6.1 Алгоритм $A - Ward$

Алгоритм A-Ward является усовершенствованием широко известного алгоритма иерархической агломеративной кластеризации Уорда (Ward)[5]. 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.

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).

7 Примеры работы с программой

7.1 Нормализация

7.1.1 Нормализация с центрированием по среднему и масштабированием по полуразмаху

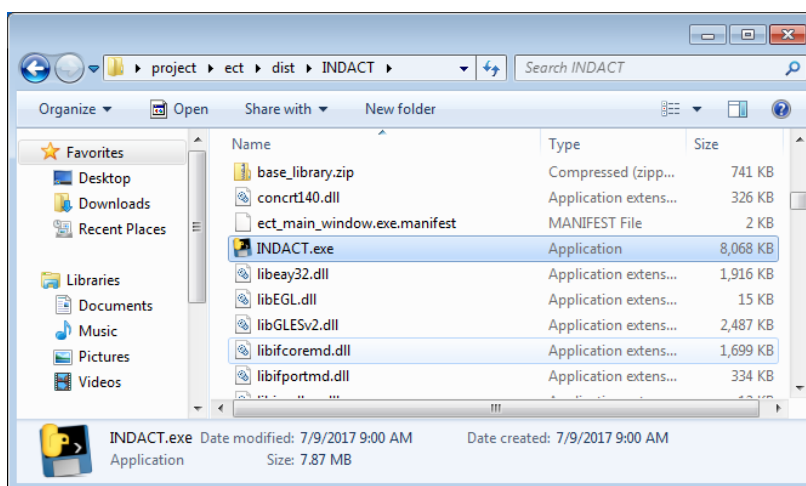
В данном разделе рассматривается пример нормализации признаков обучающего файла `smartphones.dat` с центрированием по среднему и масштабированием по полуразмаху.

Действие/Описание

Интерфейс

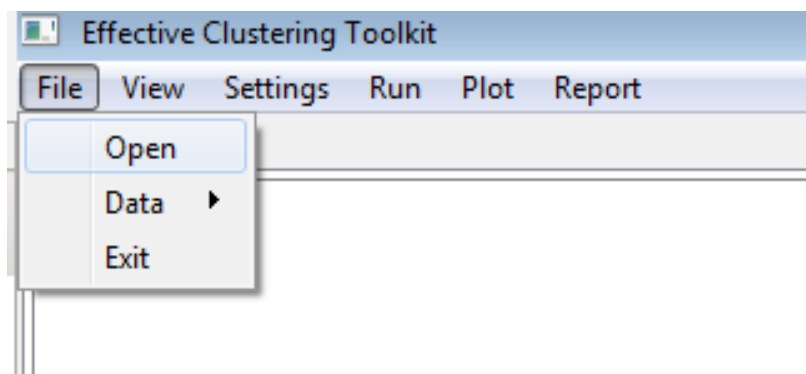
1. Запустить бинарный файл программы

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



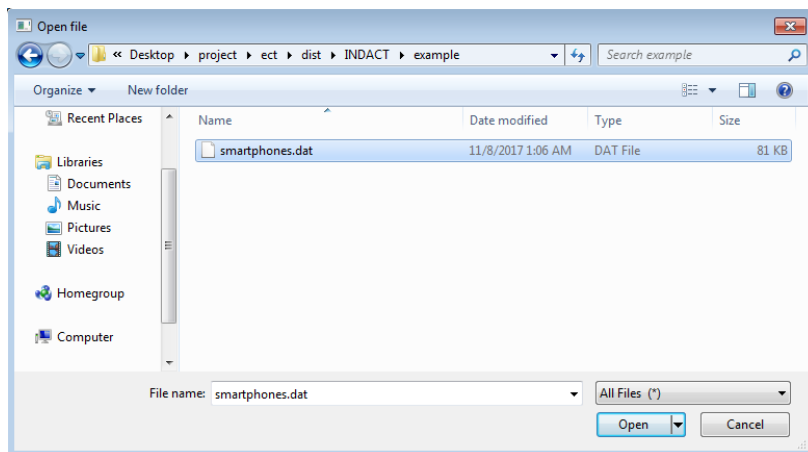
2. Открыть диалог загрузки файла

Последовательно нажать в главном меню пункты **File** ⇒ **Open**.



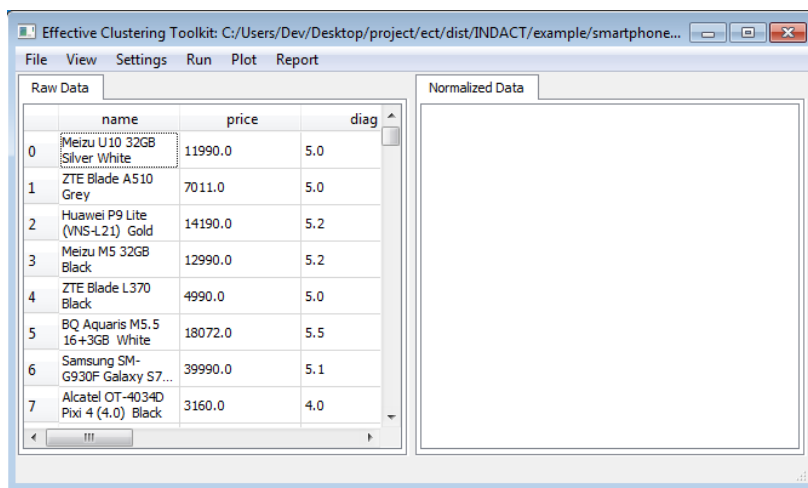
3. Выбрать текстовый файл с данными

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



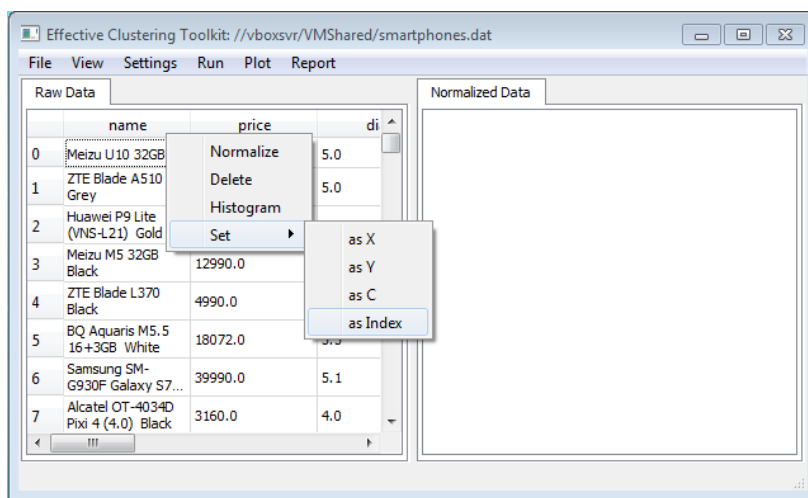
4. Убедиться что файл загружен

Посмотреть, что вкладка “Raw Data” заполнилась данными из файла.



5. Установить признак `name` как индекс

Открыть контекстное меню ПКМ и выбрать **Set** ⇒ **As Index**



6. Убедиться что признак выставился

Результат выполнения предыдущей операции показан на рисунке справа. Теперь числовые индексы заменены названиями телефонов.

	price	diag
Meizu U10 32GB	11990.0	5.0
ZTE Blade A510	7011.0	5.0
Huawei P9 Lite	14190.0	5.2
Meizu M5 32GB	12990.0	5.2
ZTE Blade L370	4990.0	5.0
BQ Aquaris M5.5	18072.0	5.5
Samsung SM-G930	39990.0	5.1
Alcatel OT-4034	3160.0	4.0

7. Удалить признак vendor

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

	ram	type	vendor
Meizu U10 32GB		IPS	Meizu
ZTE Blade A510		IPS	ZTE
Huawei P9 Lite		IPS	Huawei
Meizu M5 32GB		IPS	Meizu
ZTE Blade L370		TFT	ZTE
BQ Aquaris M5.5		IPS	BQ
Samsung SM-G930		Super AMOLED	Samsung
Alcatel OT-4034		TFT	Alcatel
Sony Xperia XA		IPS	Sony
ZTE Blade L5 PI		IPS	ZTE
Meizu Pro 6 64G		Super AMOLED	Meizu

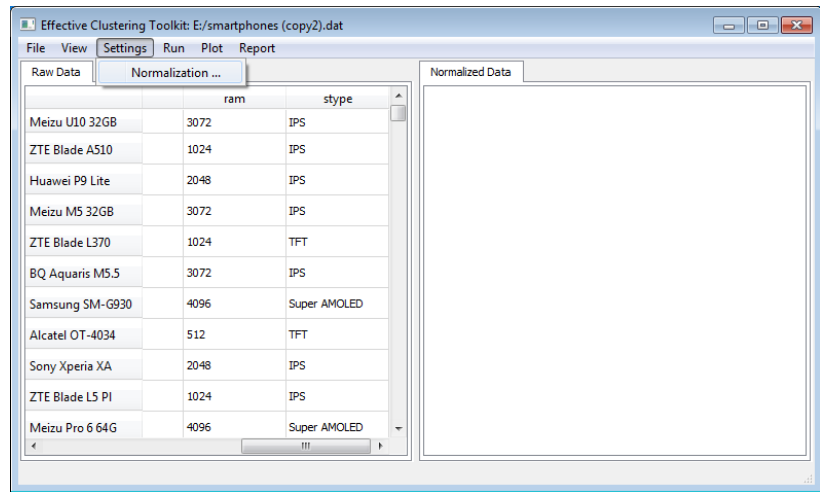
8. Убедиться что признак удалён

Результат удаления признака **vendor** показан на рисунке справа. Видно, что признак **vendor** больше не отображается во вкладке "Raw Data".

	ram	sttype
Meizu U10 32GB	3072	IPS
ZTE Blade A510	1024	IPS
Huawei P9 Lite	2048	IPS
Meizu M5 32GB	3072	IPS
ZTE Blade L370	1024	TFT
BQ Aquaris M5.5	3072	IPS
Samsung SM-G930	4096	Super AMOLED
Alcatel OT-4034	512	TFT
Sony Xperia XA	2048	IPS
ZTE Blade L5 PI	1024	IPS
Meizu Pro 6 64G	4096	Super AMOLED

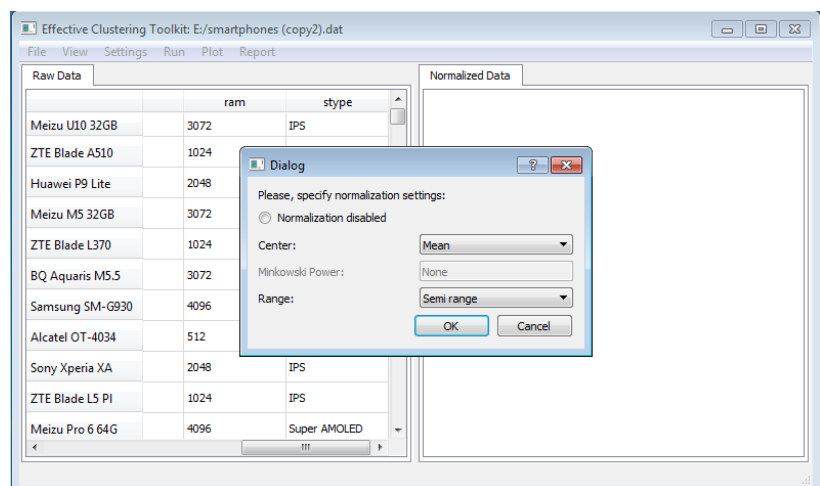
9. Открыть окно
нормализации

Выбрать в главном меню пункты **Settings** ⇒ **Normalization**.



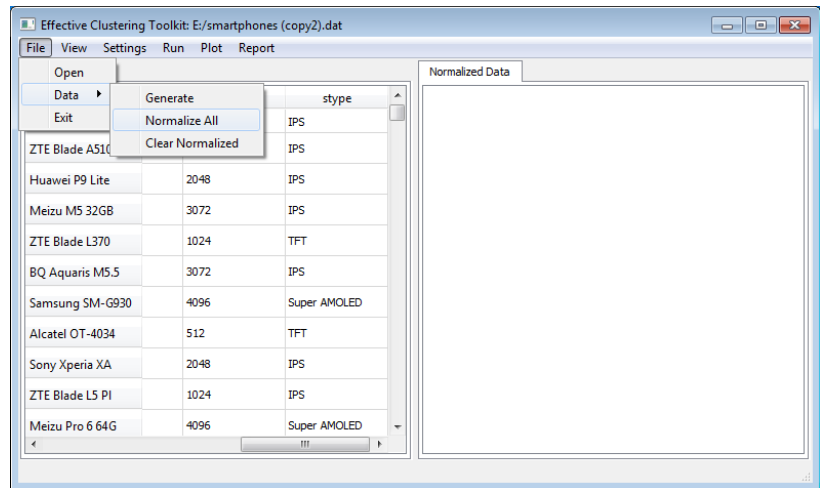
10. Выставить
параметры нормализации

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



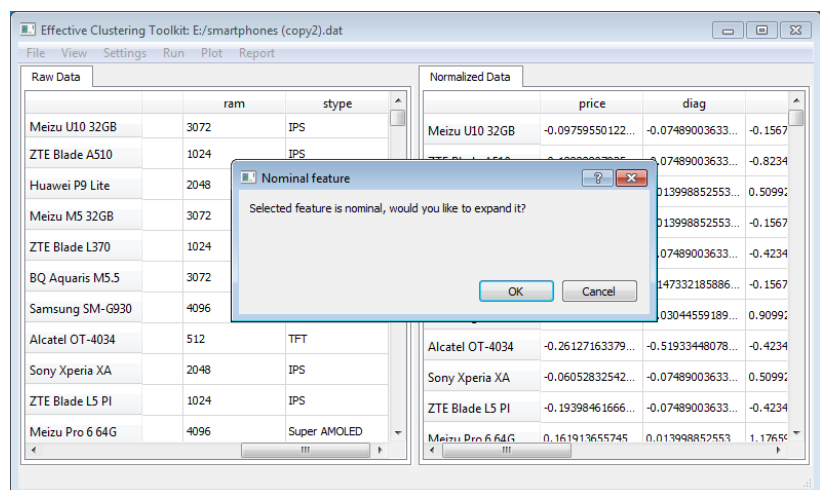
11. Запустить нормализацию всех признаков

Для запуска нормализации всех признаков сразу, требуется в главном меню выбрать **File** ⇒ **Data** ⇒ **Normalize All**.



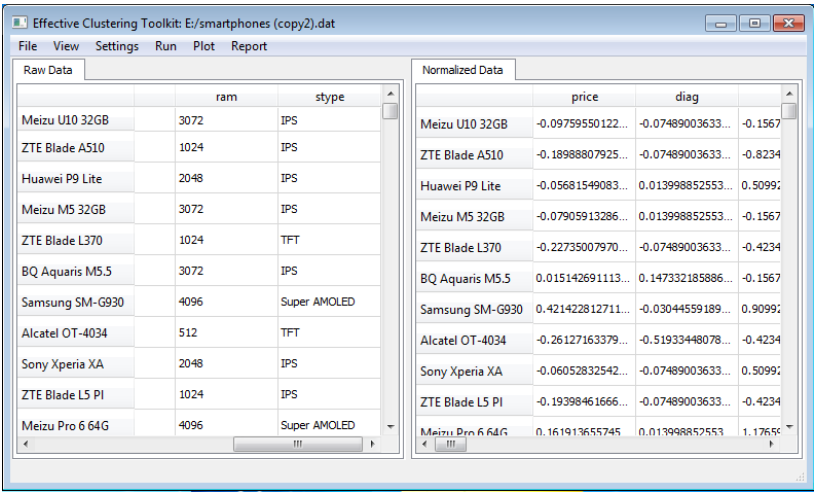
12. Подтвердить нормализацию категориального признака

Так как данные включают в себя категориальный признак **stype**, то программа запросит подтверждение разложения признака по количеству уникальных значений. Нажать кнопку **OK**.



13. Посмотреть результат

После нормализации признаков результат будет отображен во вкладке “Normalized Data”



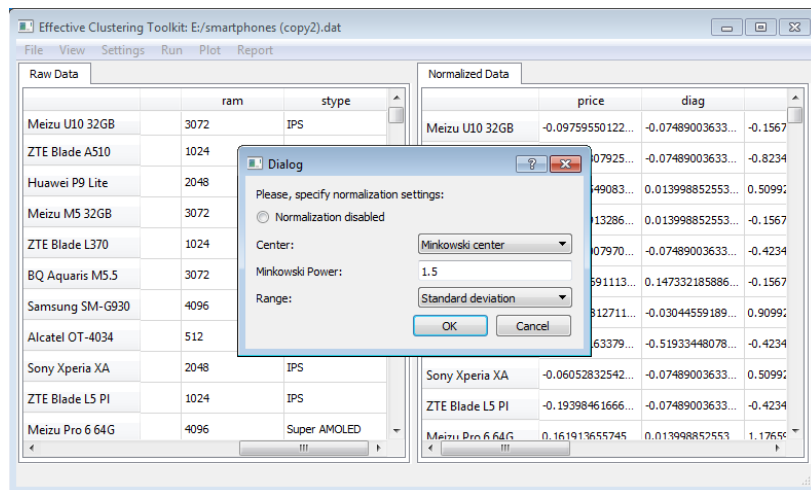
7.1.2 Нормализация с центрированием по Минковскому и масштабированием по стандартному отклонению

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

Действие/Описание	Интерфейс
1. Запустить программу и загрузить файл	Выполнить пункты 1–9 из предыдущего примера (7.1.1).

2. Выставить параметры нормализации

Выставить параметры нормализации как показано на рисунке справа. Переключатель “Normalization disabled” должен быть снят, значение Center выбрано Minkowski Center, величина Minkowski Power выставлена равной 1.5, а параметр Range выбран Standard deviation.

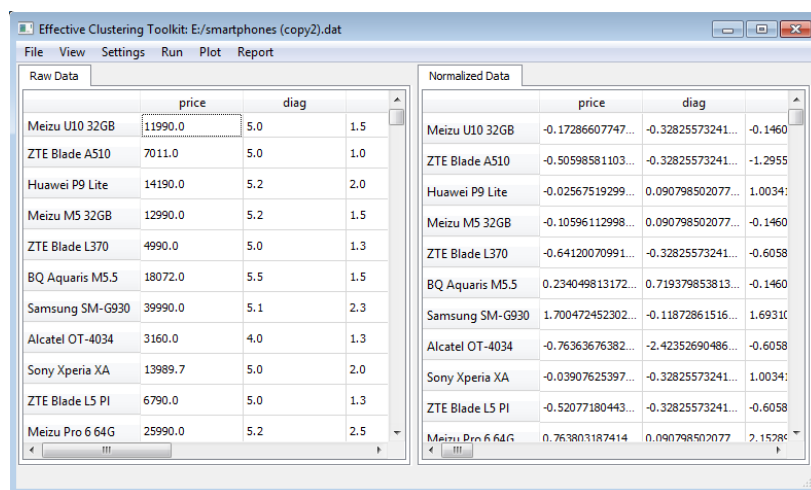


Выполнить пункты 11–12 из предыдущего примера (7.1.1).

3. Нормализовать все признаки

4. Посмотреть результат

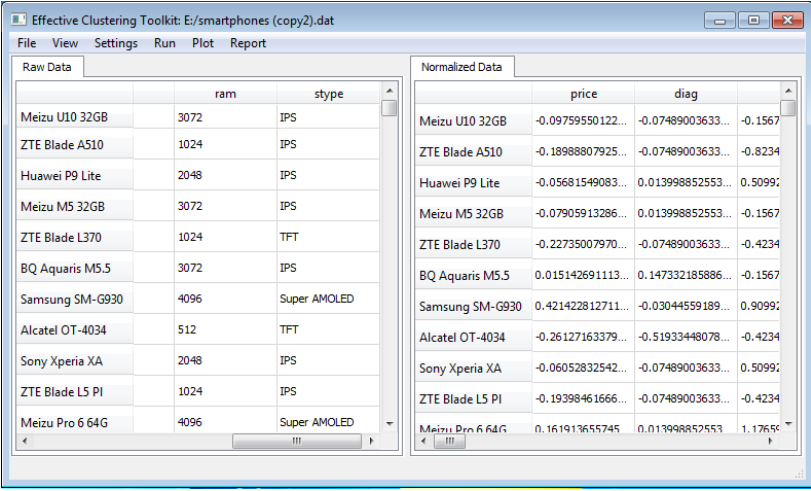
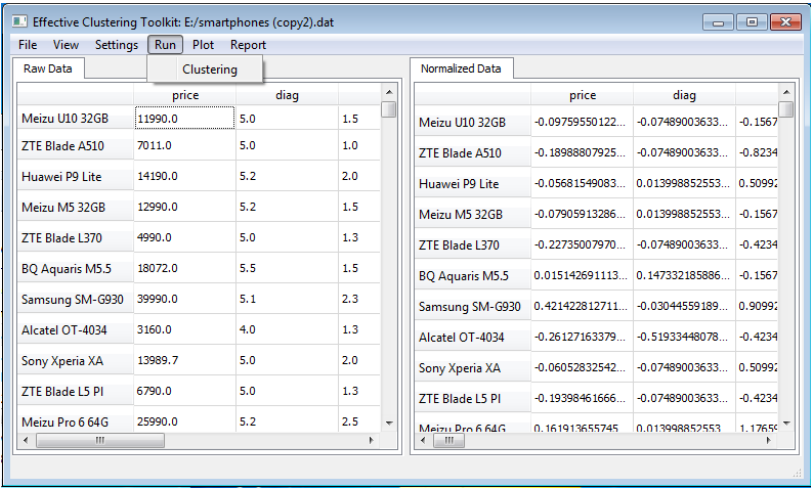
После нормализации признаков результат будет отображён во вкладке “Normalized Data”



7.2 Кластеризация

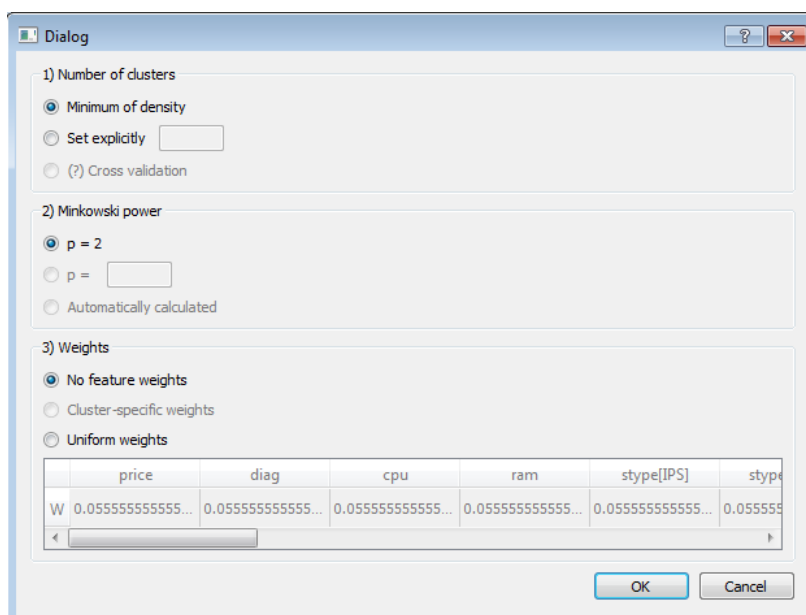
7.2.1 Кластеризация с автоматическим выбором числа кластеров

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
<p>1. Запустить программу, загрузить файл и нормализовать признаки</p> <p>Выполнить все пункты из первого примера(7.1.1). Для кластеризации требуются нормализованные признаки.</p>	
<p>2. Открыть окно кластеризации</p> <p>Пред запуском кластеризации потребуется задать общие параметры процедуры, по которым программа выберет конкретный алгоритм. Для этого необходимо из главного меню выбрать: Run ⇒ Clustering.</p>	

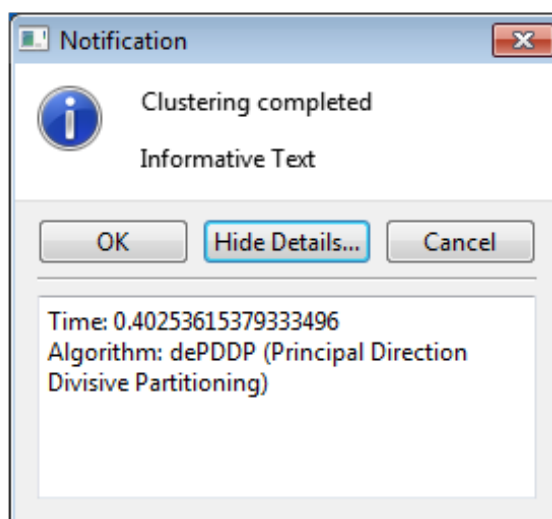
3. Установить параметры алгоритма кластеризации

Установить параметры кластеризации как показано на рисунке справа. В группе регулирующей число кластеров установить переключатель **Minimum of density** для выбора числа кластеров по минимуму функции плотности. Степень Минковского установить равной 2, отметив соответствующий переключатель. Веса признаков в данном примере не используются. Нажать **OK**.



4. Дождаться завершения кластеризации

Когда алгоритм закончит работу, появится окно, изображённое справа. Если в окне нажать кнопку **Show Details** то появится дополнительная информация о выбранном алгоритме и времени работы. Нажать кнопку **OK**.



5. Посмотреть кластерную принадлежность

После завершения кластеризации во вкладке “Normalized Data” будет заполнен столбец **Cluster#** как показано на рисунке справа.

Raw Data	price	diag	cp
Meizu U10 32GB	11990.0	5.0	1.5
ZTE Blade A510	7011.0	5.0	1.0
Huawei P9 Lite	14190.0	5.2	2.0
Meizu M5 32GB	12990.0	5.2	1.5
ZTE Blade L370	4990.0	5.0	1.3
BQ Aquaris M5.5	18072.0	5.5	1.5
Samsung SM-G930	39990.0	5.1	2.3
Alcatel OT-4034	3160.0	4.0	1.3
Sony Xperia XA	13989.7	5.0	2.0
ZTE Blade L5 PI	6790.0	5.0	1.3
Meizu Pro 6 64G	25990.0	5.2	2.5

Normalized Data	LED	stype[LCD]	Cluster#
Meizu U10 32GB	4079...	-0.00344234079...	0
ZTE Blade A510	4079...	-0.00344234079...	0
Huawei P9 Lite	4079...	-0.00344234079...	3
Meizu M5 32GB	4079...	-0.00344234079...	0
ZTE Blade L370	4079...	-0.00344234079...	1
BQ Aquaris M5.5	4079...	-0.00344234079...	0
Samsung SM-G930	4079...	-0.00344234079...	2
Alcatel OT-4034	4079...	-0.00344234079...	1
Sony Xperia XA	4079...	-0.00344234079...	3
ZTE Blade L5 PI	4079...	-0.00344234079...	0
Meizu Pro 6 64G	4079...	-0.00344234079...	2

6. Сгенерировать отчёт

Для генерации отчёта в главном меню выбрать **Report** ⇒ **Show** (см. 5.9).

Raw Data	price	diag	cp
Meizu U10 32GB	11990.0	5.0	1.5
ZTE Blade A510	7011.0	5.0	1.0
Huawei P9 Lite	14190.0	5.2	2.0
Meizu M5 32GB	12990.0	5.2	1.5
ZTE Blade L370	4990.0	5.0	1.3
BQ Aquaris M5.5	18072.0	5.5	1.5
Samsung SM-G930	39990.0	5.1	2.3
Alcatel OT-4034	3160.0	4.0	1.3
Sony Xperia XA	13989.7	5.0	2.0
ZTE Blade L5 PI	6790.0	5.0	1.3
Meizu Pro 6 64G	25990.0	5.2	2.5
BQ Aquaris X5 P	17290.0	5.0	1.8
Sony Xperia X C	24989.9	4.6	1.8
ZTE Blade V7 R	14590.0	5.2	1.3
Lenovo Vibe Sho	16890.0	5.0	1.7
HTC 10 Lifestyl	27990.0	5.2	1.8

Report	price	diag	cp
Meizu U10 32GB	-0.09759550122...	-0.07489003633...	-0.156741
ZTE Blade A510	-0.18988807925...	-0.07489003633...	-0.823407
Huawei P9 Lite	-0.05681549083...	0.013998852553...	0.509925
Meizu M5 32GB	-0.07905913286...	0.013998852553...	-0.156741
ZTE Blade L370	-0.22735007970...	-0.07489003633...	-0.423407
BQ Aquaris M5.5	0.015142691113...	0.147332185886...	-0.156741
Samsung SM-G930	0.421422812711...	-0.03044559189...	0.909925
Alcatel OT-4034	-0.26127163379...	-0.51933448078...	-0.423407
Sony Xperia XA	-0.06052832542...	-0.07489003633...	0.509925
ZTE Blade L5 PI	-0.19398461666...	-0.07489003633...	-0.423407
Meizu Pro 6 64G	0.161913655745...	0.013998852553...	1.176592
BQ Aquaris X5 P	0.000647251059...	-0.07489003633...	0.243258
Sony Xperia X C	0.143375433754...	-0.25266781411...	0.243258
ZTE Blade V7 R	-0.04940094349...	0.013998852553...	-0.423407
Lenovo Vibe Sho	-0.00676729628...	-0.07489003633...	0.109925
HTC 10 Lifestyl	0.198986392455...	0.013998852553...	0.243258

7. Посмотреть отчёт

Вид отчёта показан на рисунке справа. Сверху показана таблица интегрального представления, а снизу — текстовый отчёт.

	price	diag	cpu	ram	stype[IPS]	stype[TFT]	uper AH	[SuperL]	[Retina]	pe[LT]	type[TN]	[Super]	Super / e	e[AMOI]	[SuperL]	ype[IGZO]	ype[OLED]	ype[LCD]	Entities
0	8631.206	5.152	1.286	1645.862	1.000	0.000	-0.000	-0.000	0.000	0.000	-0.000	-0.000	-0.000	-0.000	0.000	0.000	0.000	0.000	247
1	7363.508	4.620	1.277	3372.480	-0.000	0.980	0.000	0.000	-0.000	-0.000	0.020	0.000	-0.000	0.000	-0.000	-0.000	-0.000	-0.000	50
2	28398.179	5.231	1.942	3061.989	-0.000	0.000	0.363	0.039	0.168	0.168	0.073	0.073	0.017	0.073	0.006	0.011	0.006	0.006	179
3	23255.670	5.363	2.007	3159.771	1.000	-0.000	-0.000	0.000	0.000	0.000	-0.000	-0.000	-0.000	-0.000	-0.000	-0.000	-0.000	-0.000	105
Mean	16912.141	5.091	1.628	2260.026	0.500	0.245	0.091	0.010	0.042	0.042	0.023	0.018	0.004	0.018	0.001	0.003	0.001	0.001	581

```

Intelligent K-Means resulted in 10 clusters;
Algorithm used: anomalous clustering + A-Ward
Normalization:
  center: Mean
  range: Semi range
Anomalous pattern cardinality to discard: N/A
Features involved:
  price: mean = 1.73e+04; std = 1.49e+04;
  diag: mean = 5.17; std = 0.477;
  cpu: mean = 1.62; std = 0.435;
  ram: mean = 2.32e+03; std = 1.26e+03;
  stype[IPS]: mean = 0.606; std = 0.489;
  stype[TFT]: mean = 0.0843; std = 0.278;
  stype[Super AMOLED]: mean = 0.112; std = 0.315;
  stype[SuperLCD 5]: mean = 0.012; std = 0.109;
  stype[Retina IPS]: mean = 0.0516; std = 0.221;
  stype[LTPS]: mean = 0.0516; std = 0.221;
  stype[TN]: mean = 0.0241; std = 0.153;
  stype[SuperLCD]: mean = 0.0224; std = 0.148;
  stype[HD Super AMOLED]: mean = 0.00516; std = 0.0717;
  stype[AMOLED]: mean = 0.0224; std = 0.148;
  stype[SuperLCD 3]: mean = 0.00172; std = 0.0415;
  stype[IGZO]: mean = 0.00344; std = 0.0586;
  stype[OLED]: mean = 0.00172; std = 0.0415;
  stype[LCD]: mean = 0.00172; std = 0.0415;
Cluster-specific info:
Cluster #0 [3 entities]:
  centroid (real): ['55281.300', '4.867', '1.733', '2730.667', '0.
  
```

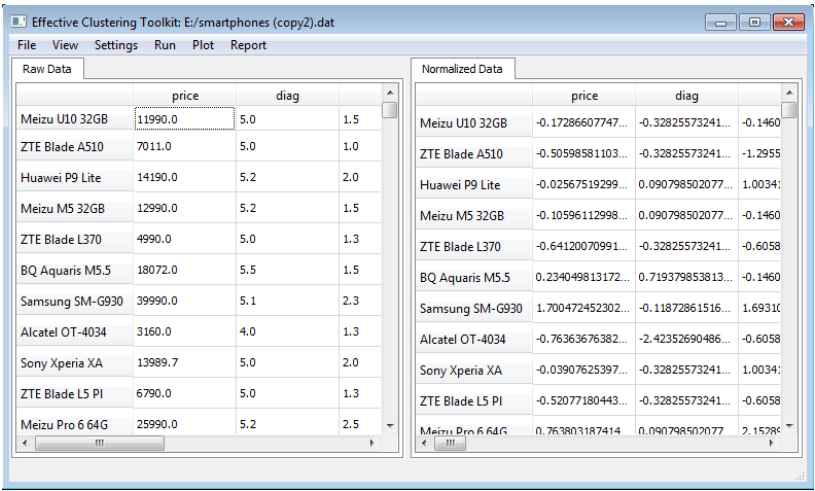
7.2.2 Кластеризация с явно заданным числом кластеров

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

Действие/Описание	Интерфейс
-------------------	-----------

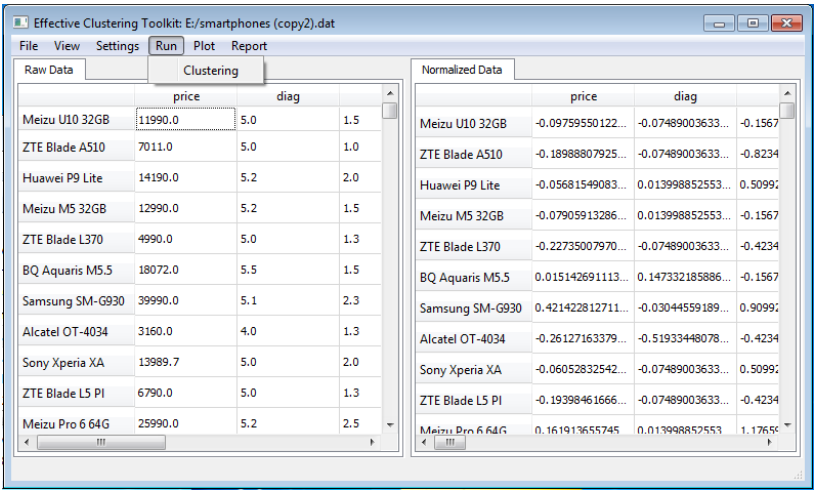
1. Запустить программу, загрузить файл и нормализовать признаки

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

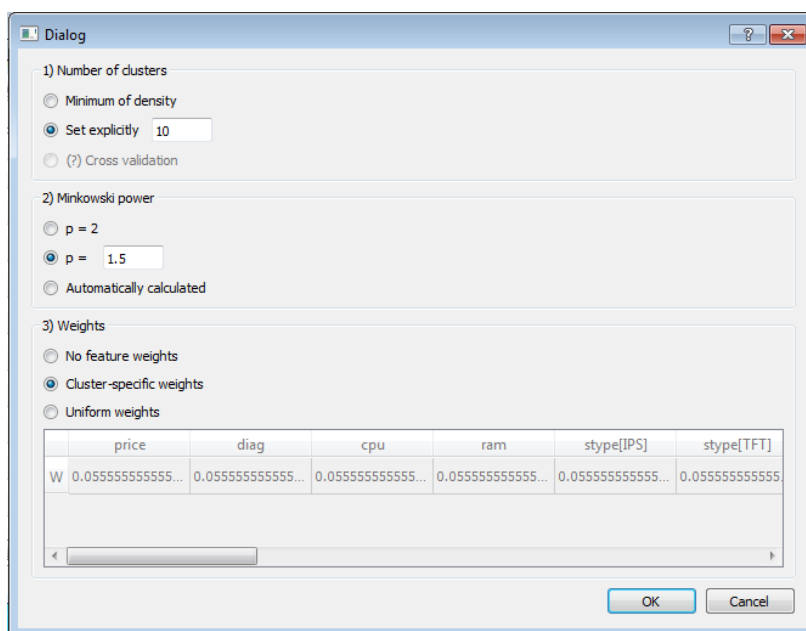


2. Открыть окно кластеризации

Пред запуском кластеризации потребуется задать общие параметры процедуры, по которым программа выберет конкретный алгоритм. Для этого необходимо из главного меню выбрать: **Run** ⇒ **Clustering**.

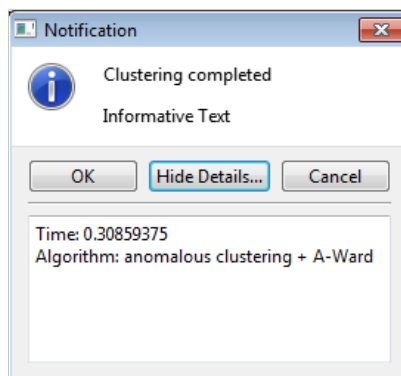


Установить параметры кластеризации как показано на рисунке справа. В группе регулирующей число кластеров установить переключатель **Set explicitly** для ввода числа кластеров с клавиатуры. Ввести число кластеров 10. Степень Минковского установить равной 1.5, отметив переключатель **p=**. Для назначения кластер-специфичных весов выбрать переключатель **Cluster-specific weights** (см. [1]). Нажать **OK**.



4. Дождаться завершения кластеризации

Когда алгоритм закончит работу, появится окно, изображённое справа. Если в окне нажать кнопку **Show Details** то появится дополнительная информация о выбранном алгоритме и времени работы. Нажать кнопку **OK**.



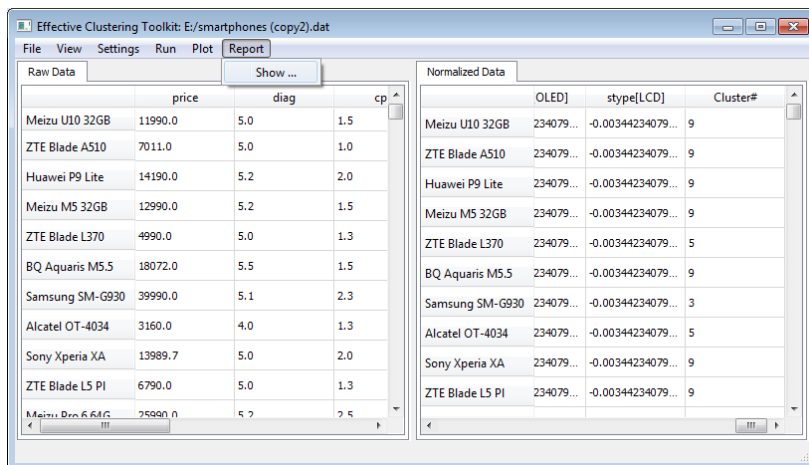
5. Посмотреть кластерную принадлежность

После завершения кластеризации во вкладке "Normalized Data" будет заполнен столбец **Cluster#** как показано на рисунке справа.

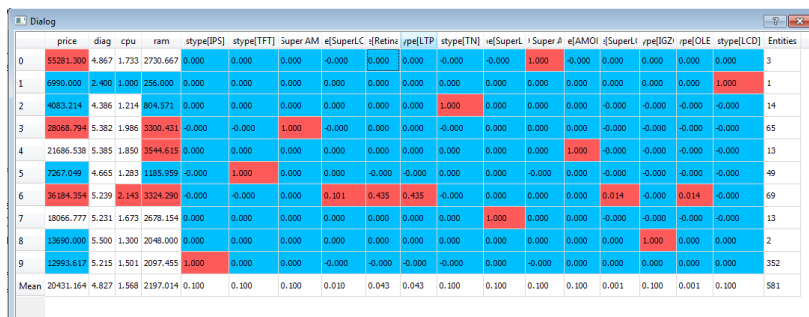
Raw Data	price	diag	cp
Meizu U10 32GB	11990.0	5.0	1.5
ZTE Blade A510	7011.0	5.0	1.0
Huawei P9 Lite	14190.0	5.2	2.0
Meizu M5 32GB	12990.0	5.2	1.5
ZTE Blade L370	4990.0	5.0	1.3
BQ Aquaris M5.5	18072.0	5.5	1.5
Samsung SM-G930	39990.0	5.1	2.3
Alcatel OT-4034	3160.0	4.0	1.3
Sony Xperia XA	13989.7	5.0	2.0
ZTE Blade L5 PI	6790.0	5.0	1.3
Meizu Pro 5 64G	25990.0	5.2	2.5

Normalized Data	OLEID	stype[LCD]	Cluster#
Meizu U10 32GB	234079...	-0.00344234079...	9
ZTE Blade A510	234079...	-0.00344234079...	9
Huawei P9 Lite	234079...	-0.00344234079...	9
Meizu M5 32GB	234079...	-0.00344234079...	9
ZTE Blade L370	234079...	-0.00344234079...	5
BQ Aquaris M5.5	234079...	-0.00344234079...	9
Samsung SM-G930	234079...	-0.00344234079...	3
Alcatel OT-4034	234079...	-0.00344234079...	5
Sony Xperia XA	234079...	-0.00344234079...	9
ZTE Blade L5 PI	234079...	-0.00344234079...	9

Для генерации отчёта в главном меню выбрать **Report** ⇒ **Show** (см. 5.9).



Вид отчёта показан на рисунке справа. Сверху показана таблица интегрального представления, а снизу — текстовый отчёт. Как видно из интегрального отчёта, некоторые кластеры не представительны, поэтому в реальных условиях, возможно, следовало бы рассмотреть другие значения параметров.



```
Report
Intelligent K-Means resulted in 10 clusters;
Algorithm used: anomalous clustering + A-Ward
Normalization:
  center: Mean
  range: Semi range
Anomalous pattern cardinality to discard: N/A
Features involved:
  price: mean = 1.73e+04; std = 1.49e+04;
  diag: mean = 5.17; std = 0.477;
  cpu: mean = 1.62; std = 0.435;
  ram: mean = 2.32e+03; std = 1.26e+03;
  stype[IPS]: mean = 0.606; std = 0.489;
  stype[TFT]: mean = 0.0843; std = 0.278;
  stype[Super AMOLED]: mean = 0.112; std = 0.315;
  stype[SuperLCD 5]: mean = 0.012; std = 0.109;
  stype[Retina IPS]: mean = 0.0516; std = 0.221;
  stype[LTPS]: mean = 0.0516; std = 0.221;
  stype[TN]: mean = 0.0241; std = 0.153;
  stype[SuperLCD]: mean = 0.0224; std = 0.148;
  stype[HD Super AMOLED]: mean = 0.00516; std = 0.0717;
  stype[AMOLED]: mean = 0.0224; std = 0.148;
  stype[SuperLCD 3]: mean = 0.00172; std = 0.0415;
  stype[IGZO]: mean = 0.00344; std = 0.0586;
  stype[OLED]: mean = 0.00172; std = 0.0415;
  stype[LCD]: mean = 0.00172; std = 0.0415;

Cluster-specific info:
Cluster #0 [3 entities]:
  centroid (real): ['55281.300', '4.867', '1.733', '2730.667', '0.0
```

Аббревиатуры

INDUCT	Intelligent Data Clustering Toolkit.
SVD	Singular Value Decomposition.
ЛКМ	Левая Кнопка Мыши.
ОЗУ	Оперативное Запоминающее Устройство.
ОС	Операционная Система.
ПК	Персональный Компьютер.
ПКМ	Правая Кнопка Мыши.
ПО	Программное Обеспечение.
СИК	Система Интеллектуальной Кластеризации.

Словарь терминов

dll библиотека динамически подключаемая библиотека позволяющая многократное использование различными программными приложениями (англ. *Dynamic Link Library*). Примером динамически подключаемой библиотеки может служить `kernel32.dll`, реализующая основные функции MS Windows, такие как управление памятью, вводом-выводом и т.д.

Python язык программирования высокого уровня на котором написана программа INDACT.

scatter plot способ графического отображения двумерных данных на плоскости при котором каждый объект соответствует точке с координатами, равными значениям признаков этого объекта.

главное меню элемент графического интерфейса программы, содержащий основные действия. Главное меню представляет собой строку в верхней части основного окна программы, отмечено цифрой 1 на рис. ??.

диапазон нормирования величина в знаменателе формулы 1.

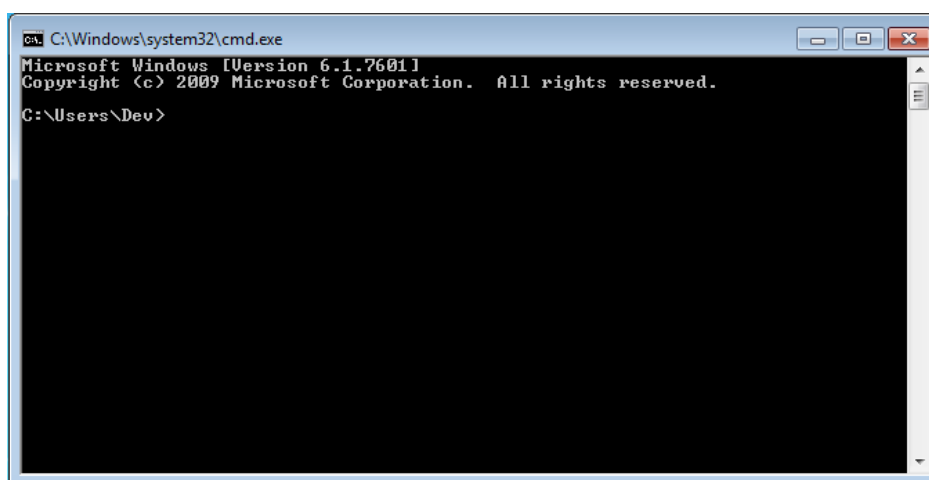
исходные данные набор данных, представленных в виде плоской таблицы и сохранённой в текстовом файле. Строка таблицы соответствует одному объекту, а столбец — признаку. Пример таблицы исходных данных показан в разделе 4.3.

каталог бинарных файлов директория файловой системы, в которой находятся исполняемые файлы программы.

кластер множество объектов исходных данных, которые обладают общими признаками и выделяются этими признаками среди остальных объектов. Например, в случае кластеризации животных в один кластер могут быть выделены животные, принадлежащие одной таксономической единице (допустим, биологическому виду).

кластер-анализ совокупность методов, разделяющих объекты таблицы наблюдений в множества (кластеры) таким образом, чтобы сходные объекты попадали в один и тот же кластер, а несходные — в разные кластеры [2]. Наиболее популярный метод кластер-анализа — k-menas.

консольное окно окно терминала Windows. Типичный вид консольного окна показан ниже:



метка вспомогательный символ, присваиваемый пользователем для определённого признака. Метки используются для выбора роли признака при построении диаграмм (например, scatter plot). 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.

нормализация данных преобразование данных с целью приведения всех признаков к одному масштабу и началу отчёта.

объект сущность предметной области, соответствующая строке в таблице данных. Например, объектом может быть определённая модель смартфона, обладающая признаками: частота процессора, диагональ экрана и т.д..

основное окно программы окно Windows, которое открывается сразу после запуска программы, см. рис. ??.

признак числовая или категориальная характеристика объекта, соответствующая столбцу в таблице данных. Например, признаками объекта “смартфон” могут быть: частота процессора, диагональ и тип экрана и т.д..

текстовый файл компьютерный файл, содержащий текстовые данные. Такой файл может быть отредактирован любым текстовым редактором, при этом разрешение файла не имеет значения (например, текстовым может быть файл *.txt или *.csv) .

утилита вспомогательная компьютерная программа для выполнения специализированных типовых задач, связанных с работой оборудования и операционной системы.

Например, для преобразования скриптов на языке Python в исполняемые `exe` файлы можно использовать утилиту `pyinstaller` (см. <http://www.pyinstaller.org/>).

центр нормирования величина вычитаемая из исходных данных в числителе формулы 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] Boris Mirkin Mikhail Tokmakov. Capturing the right number of clusters with K-Means using the complementary criterion and affinity propagation.
- [4] 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.
- [5] Joe H. Ward J. Hierarchical Grouping to Optimize an Objective Function // Journal of American Statistical Association. 1963.
- [6] G.H. Ball D.J. Hall. A clustering technique for summarizing multivariate data // Behav. Sci. 1967. С. 153–155.
- [7] Boley D. Principal Direction Divisive Partitioning // Data Mining and Knowledge Discovery. 1998. С. 325–344.
- [8] Tasoulis S.K. Tasoulis D.K. Plagianakos V.P. Enhancing Principal Direction Divisive Clustering // Pattern Recognition. 2010. С. 3391–3411.
- [9] Boley D. Principal Direction Divisive Partitioning // Data Mining and Knowledge Discovery. 1998. № 02. С. 325–344.
- [10] Tasoulis S.K. Tasoulis D.K. Plagianakos V.P. Enhancing Principal Direction Divisive Clustering // Pattern Recognition. 2010. № 43. С. 3391–3411.
- [11] Mirkin B. Core Concepts in Data Analysis: Summarization, Correlation, Visualization.