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 2017

Аннотация

Система интеллектуальной кластеризации данных 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 кластер-анализа — *k-means*. 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-means*. 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

1.1 Область применения

Программное обеспечение “Система Интеллектуальной Кластеризации” (СИК) применяется для проведения кластер-анализа таблиц данных с использованием интеллектуальных алгоритмов. Типичный пример задачи для решения который 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 СИК 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).

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 СИК является повышение эффективности анализа данных. Функционалом системы предусмотрено два типа работ:

- кластеризация реальных данных
- проведение численного эксперимента с синтетическими данными

3 Условия применения

Программный продукт работает в операционной системе Microsoft Windows 7 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

Наличие дополнительного оборудования не требуется.

4 Подготовка системы к работе и запуск

Для подготовки системы к работе требуется скопировать каталог с бинарными файлы 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.

Внимание! Запуск графического интерфейса может потребовать от 10 секунд и более в зависимости от производительности ПК пользователя.[1] Дополнительные действия не требуются.

[1] Большое время запуска графического интерфейса связано с инициализацией среды выполнения Python и загрузкой dll библиотек. В настоящее время распространение ПО для ОС Windows осуществляется при помощи утилиты `pyinstaller` (см. http://www.pyinstaller.org/), что увеличивает время запуска.

5 Основные прицины работы

5.1 Этапы работы с программой

Работа с программой СИК 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 и кластеризации, а также отбор признаков, участвующих в кластеризации и выбор основных свойств применяемого алгоритма (см. 5.4, 5.6.1). После выбора необходимых параметров пользователь запускает алгоритм кластеризации. Когда выполнение кластеризации заканчивается, пользователю становятся доступны результаты работы для просмотра и анализа.

5.2 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
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 “5.3 Обучающий файл”.

5.3 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 `example`. 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 сматрфонов, 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 "5.2 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 `data.csv`

```
          name,      price,   diag,    cpu,    ram,   stype,   vendor
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
```

5.4 Нормализация

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

$$X' = \frac{X - c}{r}, \tag{1}$$

где X — исходные данные,
c — параметр, определяющий преобразование начала отсчёта,
r — параметр, определяющий преобразование масштаба шкалы.

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.

5.5 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 таблице или в виде текстового отчёта.

5.6 Общие сведения о пользовательском интерфейсе

5.6.1 Главное окно

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

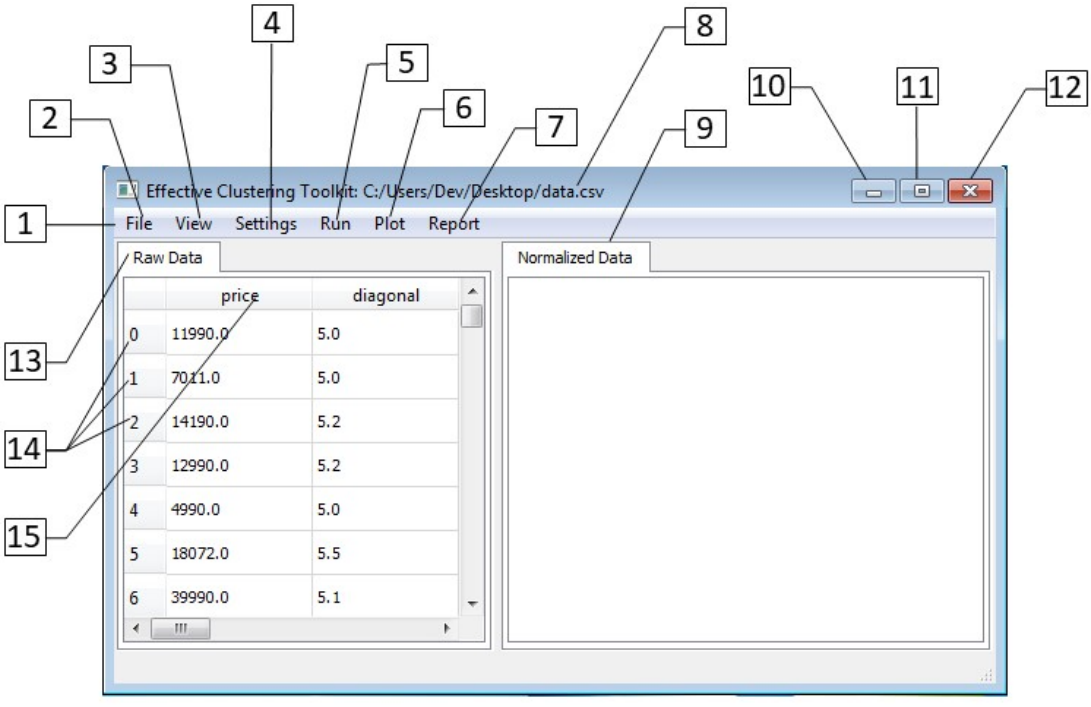


Рисунок 1 – Основные элементы пользовательского интерфейса

1. Главное меню, элемент интерфейса, содержащий основные команды
2. Меню File, содержит пункты:
   - Open — для загрузки файла данных

- Data — для манипулирования данными
- Exit — для выхода из программы

3. Меню View, содержит пункты:
   - Layout — для настройки способа отображения панелей данных: Tab Layout для отображения по вкладкам или Panal Layout для отбражения на двух панелях одновременно (по умолчанию, как показано на схеме)
4. Меню Settings, служит для настройки параметров, содержит пункты:
   - Normalization — для задания диапазона и центра нормализации
5. Меню Run, содержит пункты:
   - Clustering — для настройки и запуска алгоритма кластеризации
6. Меню Plot, служит для вызова команд графического отображения, содержит пункты:
   - Plot Data by Markers — для построения поля рассеяния (scatter plot) по отмеченным признакам
   - Delete All Markers — для удаления всех отметок признаков
   - SVD — для построения SVD диаграммы
7. Меню Report для формирования отчёта, содержит пункты:
   - Show — для отображения отчёта
8. Заголовок окна, содержит путь к открытому файлу
9. Вкладка с нормализованными данными
10. Кнопка “Свернуть окно”
11. Кнопка “Развернуть окно”
12. Кнопка “Закрыть окно”
13. Вкладка с исходными данными
14. Номера/названия объектов
15. Названия признаков

5.6.2 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`.

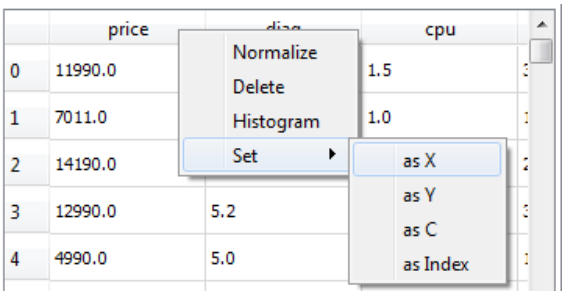


Рисунок 2 – Контекстное меню признака `price`

Контекстное меню содержит следующие пункты:

1. Normalize — нормализует выбранный признак, добавляя его во вкладку “`Normalized Data`” (см. 6.3.2)
2. Delete — удаляет признак из вкладки, в которой вызвано контекстное меню (см. 6.4.1)
3. Histogram — строит гистограмму по выбранному признаку (см. 6.6.1)
4. Set — устанавливает особые свойства для признака (см. 6.6.2, 6.4.3)
   4.1. as X — выставить метку `X` для признака (см. 6.6.2)
   4.2. as Y — выставить метку `Y` для признака (см. 6.6.2)
   4.3. as C — выставить метку `C` для признака (см. 6.6.2)
   4.4. as Index — установить признак как индекс (см. 6.4.3)

5.6.3 Диалог нормализации

На рисунке 3 показан диалог нормализации. Это окно требует от пользователя выставить значения для проведения нормализации (см. 6.3).

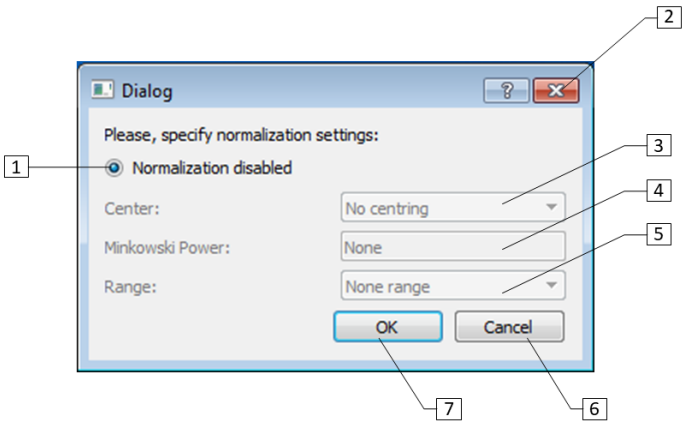


Рисунок 3 – Диалог установки параметров нормализации

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

5.6.4 Окно графической информации

Окно графической информации служит для просмотра различного вида графиков и диаграмм. Такое окно может встретиться пользователю, например при построении гистограммы (раздел 6.6.1), SVD диаграммы (6.6.3) или поля рассеяния (6.6.2).

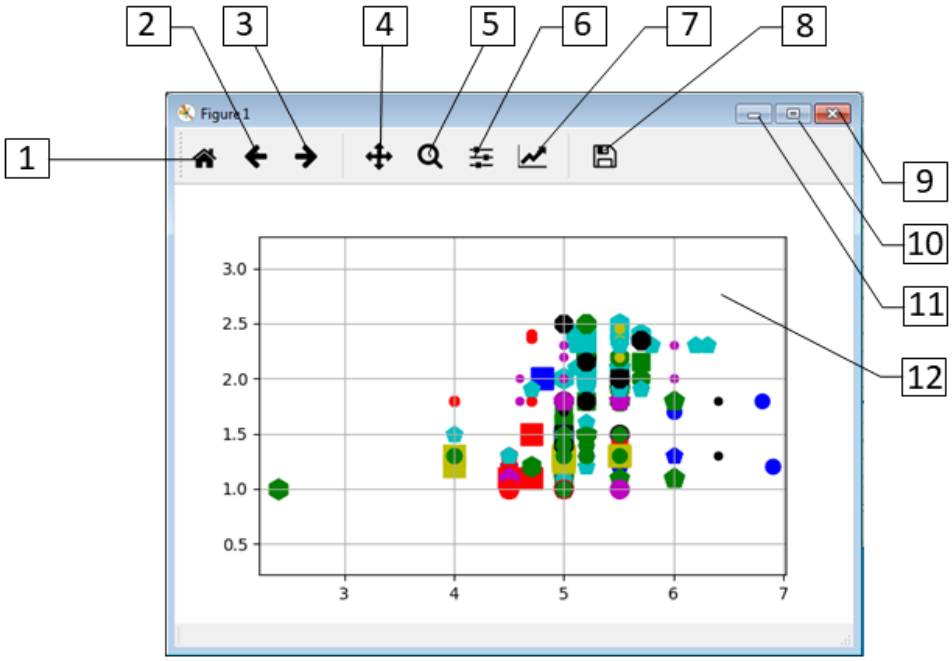


Рисунок 4 – Окно графической информации

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 конфигурации отображения
7. Кнопка редактирования осей
8. Кнопка сохранения текущего графика в файл

5.6.5 Окно генерации данных

Окно генерации применяется при работе с синтетическими данными (см. раздел 6.7). Это окно необходимо для ввода информации о значениях характеристик генерируемых данных.

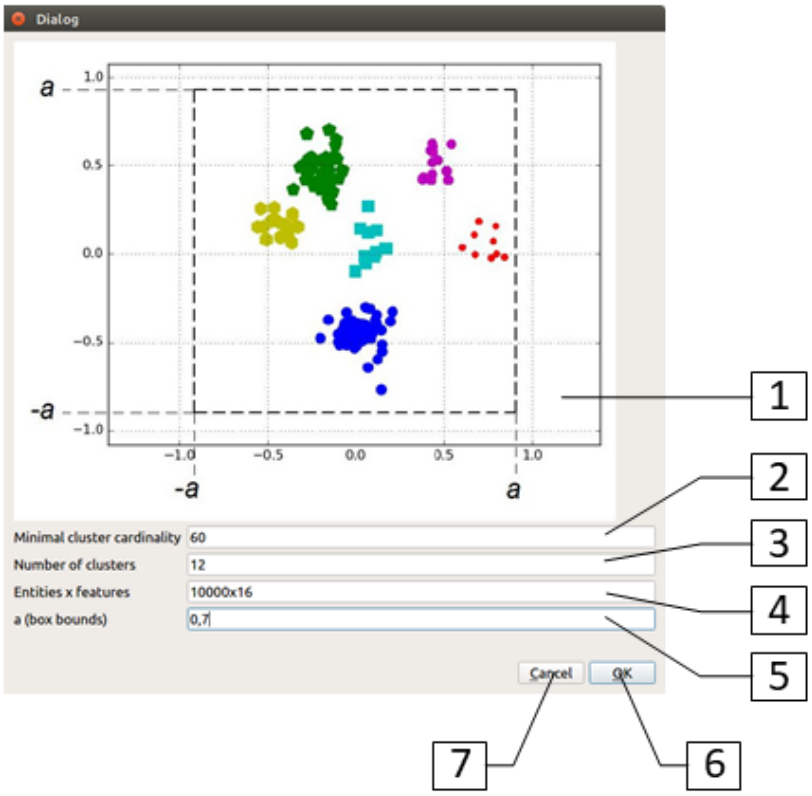


Рисунок 5 – Окно генерации данных

1. Графическая иллюстрация смысла параметра a
2. Поле ввода минимального числа объектов в каждом кластере
3. Поле ввода числа кластеров
4. Поле ввода размерности данных (число объектов `x` число признаков)
5. Поле ввода параметра a
6. Кнока подтверждения ввода

7. Кнопка отмены ввода

5.6.6 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.

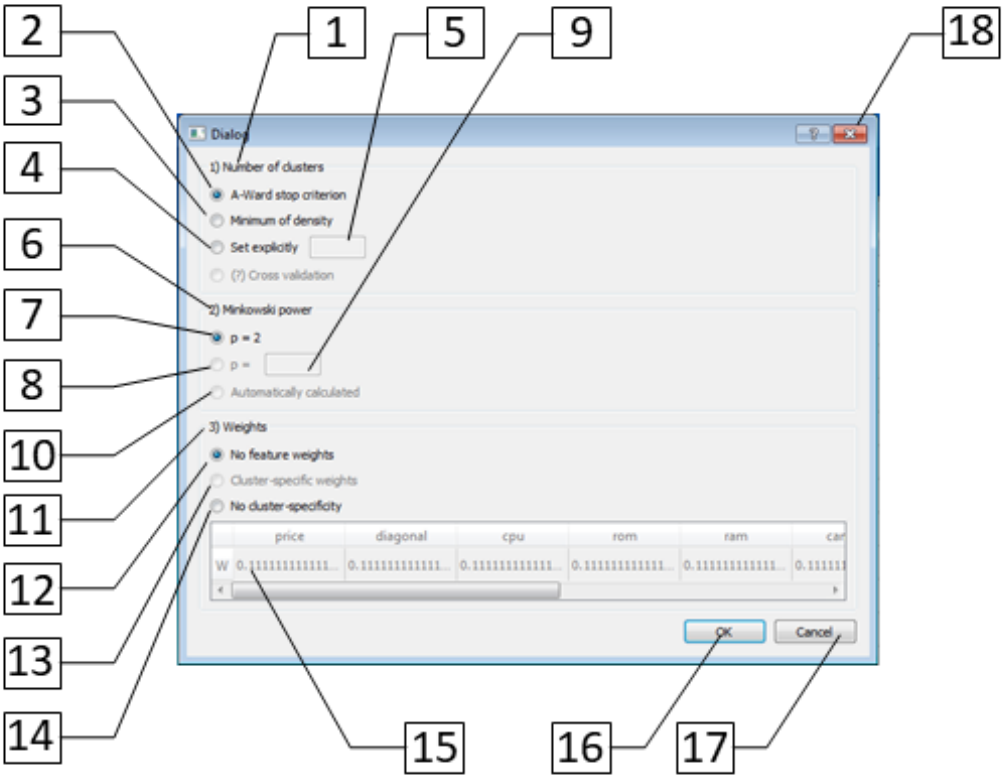


Рисунок 6 – Окно запуска кластеризации

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 задание 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 будующих версиях)
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

5.6.7 Окно таблицы результатов

Окно таблицы результатов служит для интегрального представления полученной кластерной структуры. Это окно может быть вызвано после выполнения шага кластеризации (см. 6.9). 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.

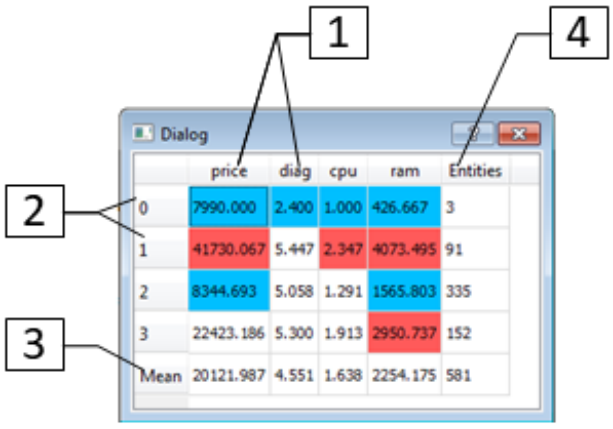


Рисунок 7 – Окно таблицы результатов

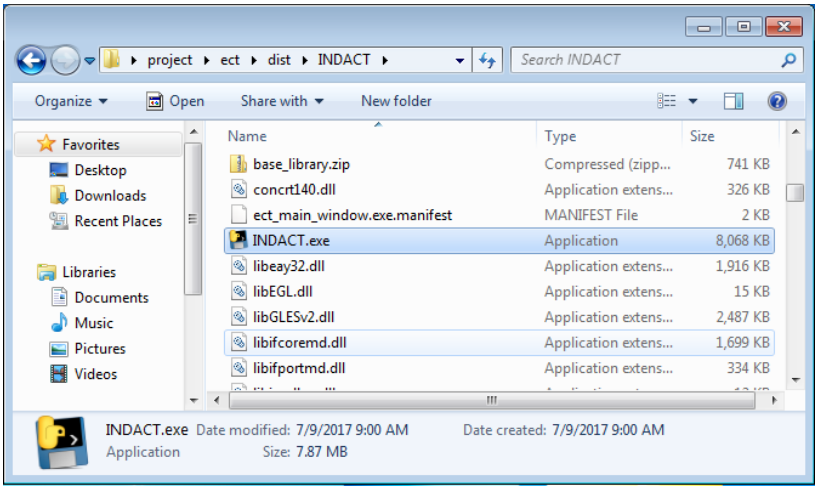
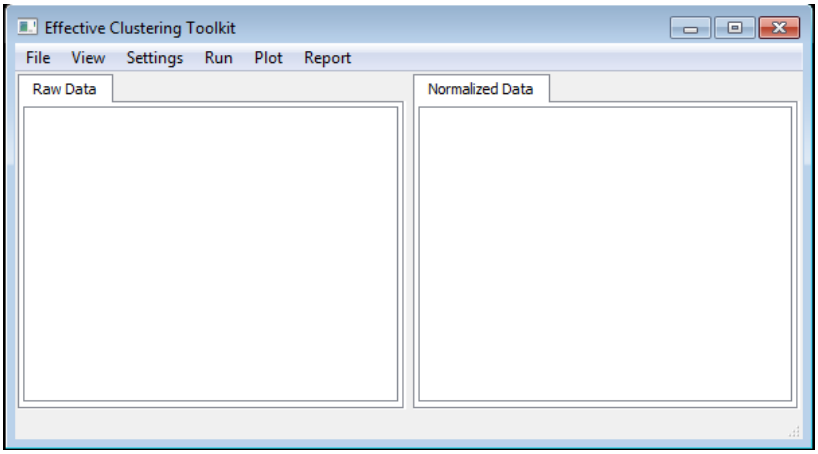
1. Название признаков

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

6 Описание операций

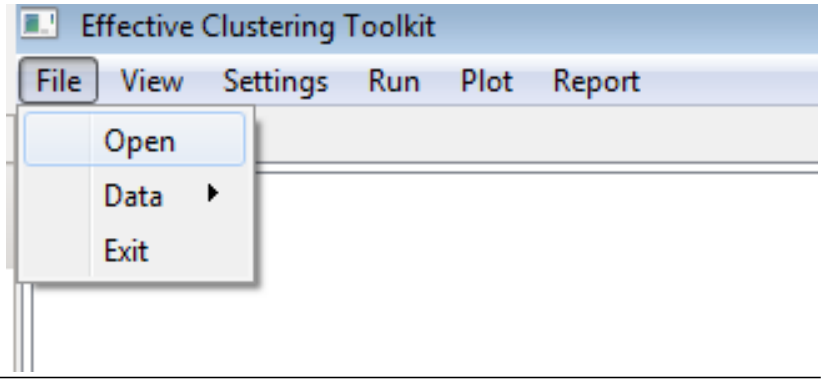
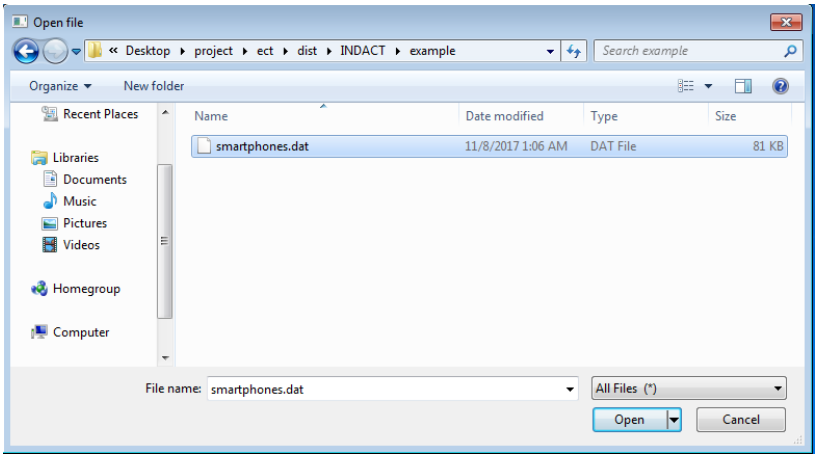
6.1 Запуск программы

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
1. Запустить бинарный файл программы Дважды нажать левой кнопкой мыши (ЛКМ) на значке `INDACT.exe`	
2. Дождаться запуска Подождать, пока произойдёт инициализация среды выполнения Python. Открытие чёрного консольного окна, означает что установлена отладочная версия программы. Его не следует закрывать.	

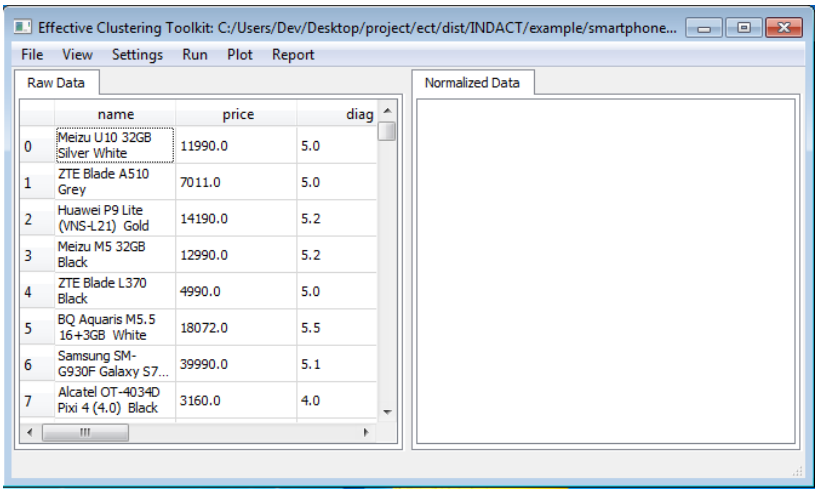
6.2 Загрузка исходных данных

Загрузка данных необходима для того чтобы подать программе файл, который содержит таблицу данных. Формат файла должен удовлетворять набору требований, перечисленных в разделе “5.2 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

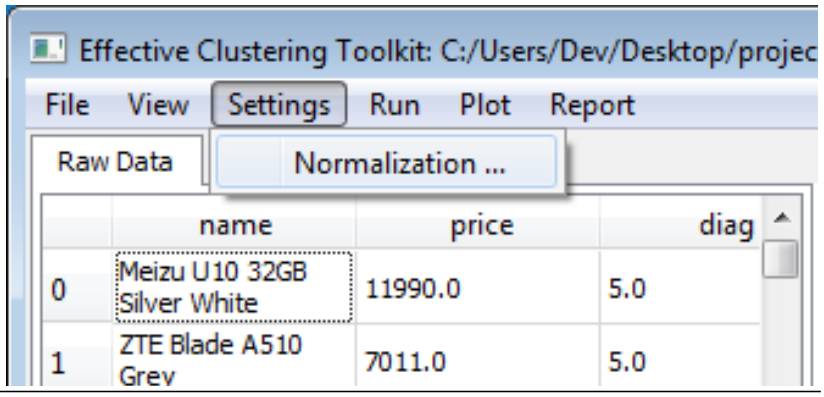
После выполнения предыдущего пункта будет произведена загрузка файла и отображение его содержимого в виде таблицы в интерфейсе программы. Пользователю следует убедиться, что загружен правильный файл, объекты и признаки отображаются верно. На рисунке справа показан загруженный файл `smartphones.dat`



6.3 Нормализация

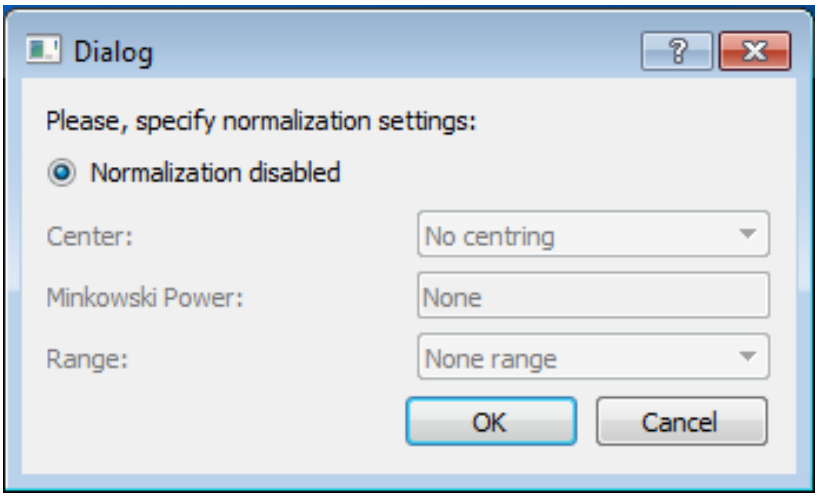
6.3.1 Установка параметров нормализации

Назначение и параметры нормализации описаны в разделе “5.4 Нормализация”.

Действие/Описание	Интерфейс
1. Открыть диалог нормализации Диалог нормализации можно открыть выбрав в главном меню пункты `Settings` ⇒ `Normalization`.	

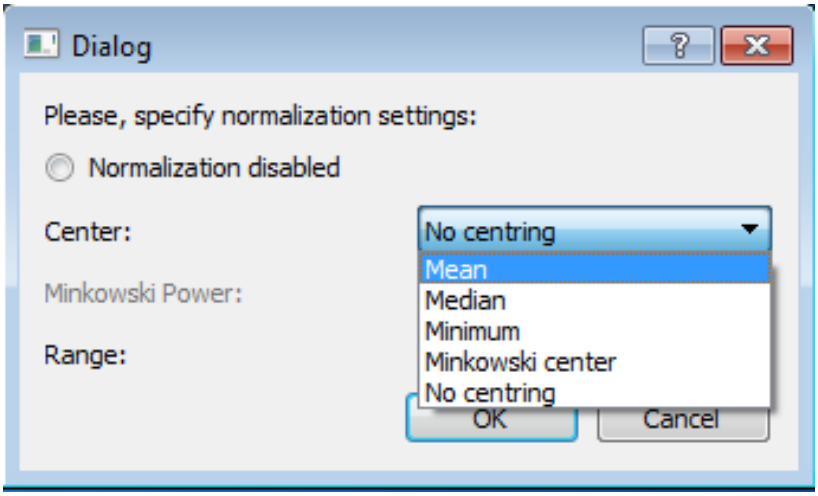
2. Включить нормализацию

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



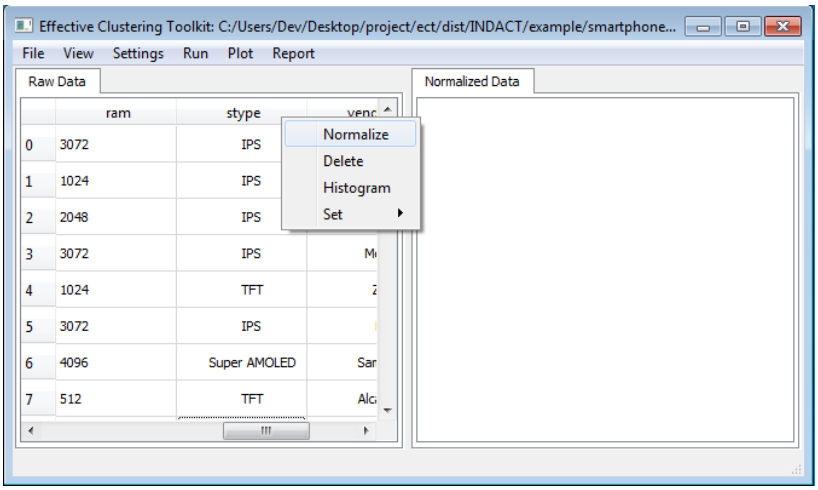
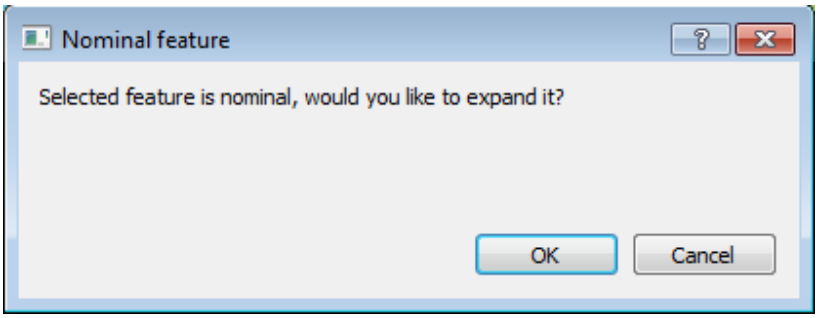
3. Выставить параметры

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 8.1.1.

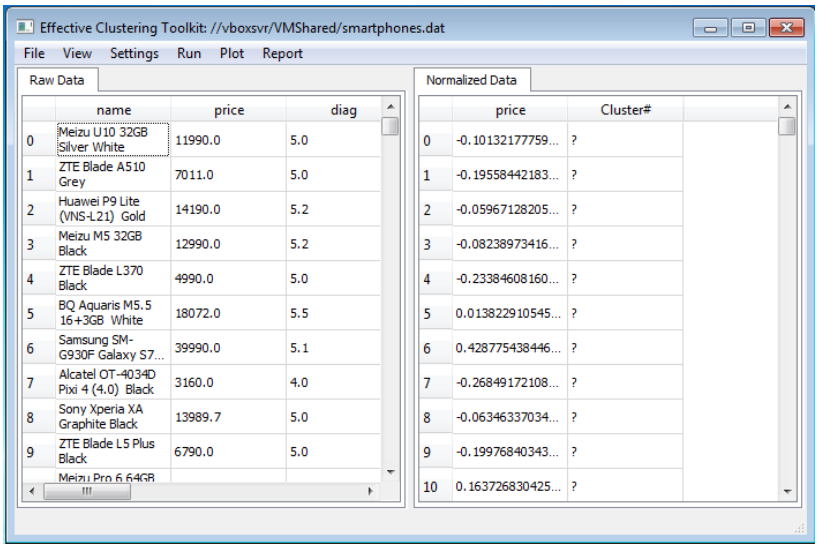


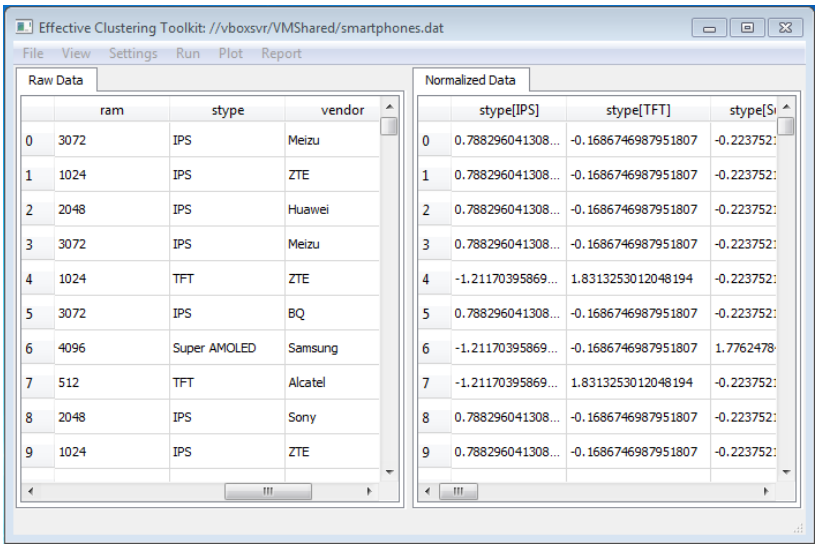
6.3.2 Нормализация одного признака

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

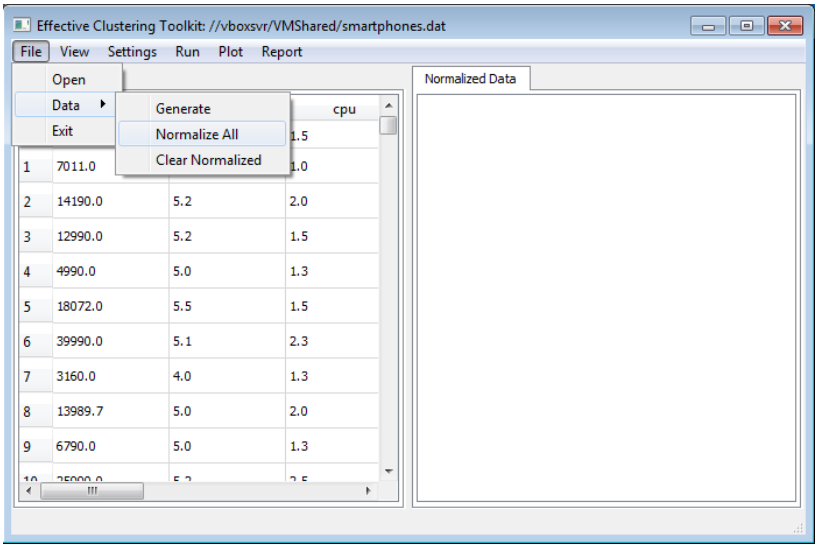
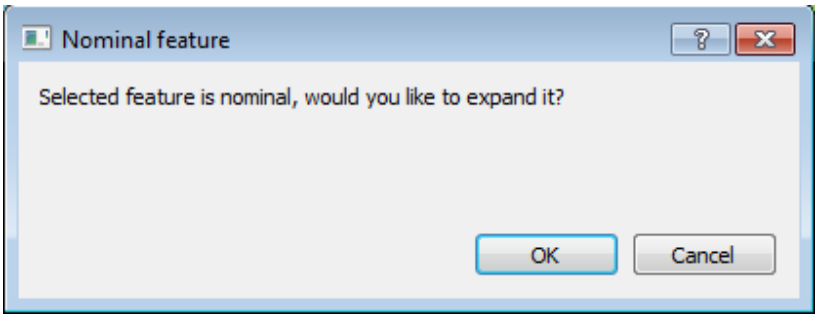




После выбора признака, он будет перенесён из вкладки “`Raw Data`” во вкладку “`Normalized Data`” и к нему будут применены выбранные настройки нормализации. Дополнительно во вкладке “`Normalized Data`” будет отображён столбец “`Cluster#`”, который будет оставаться заполненным символами “?” до тех пор, 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)

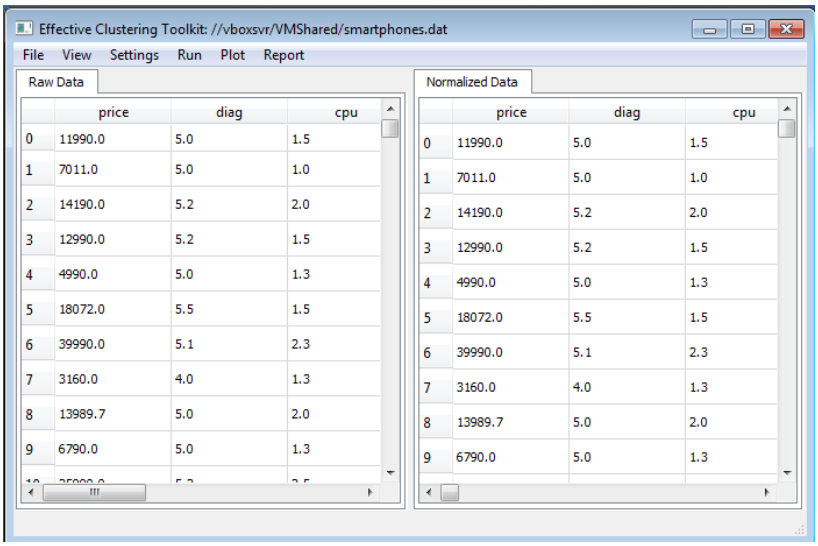
6.3.3 Нормализация всех признаков сразу

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

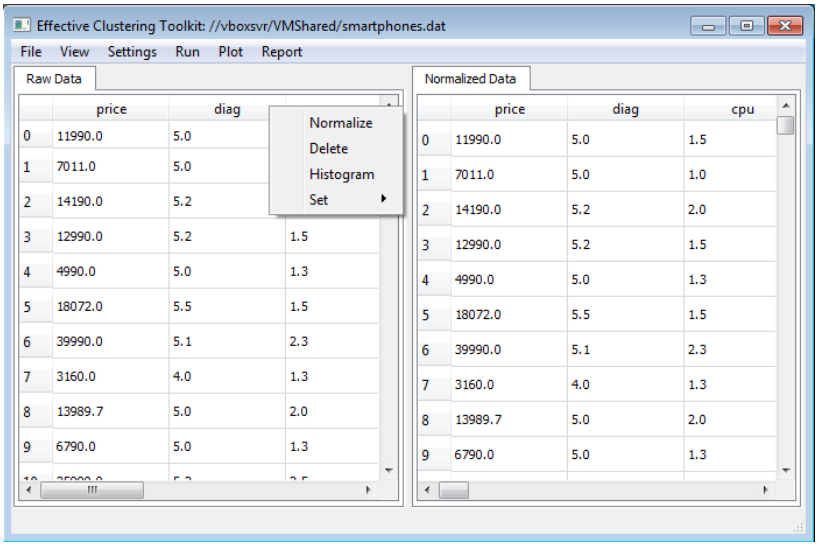
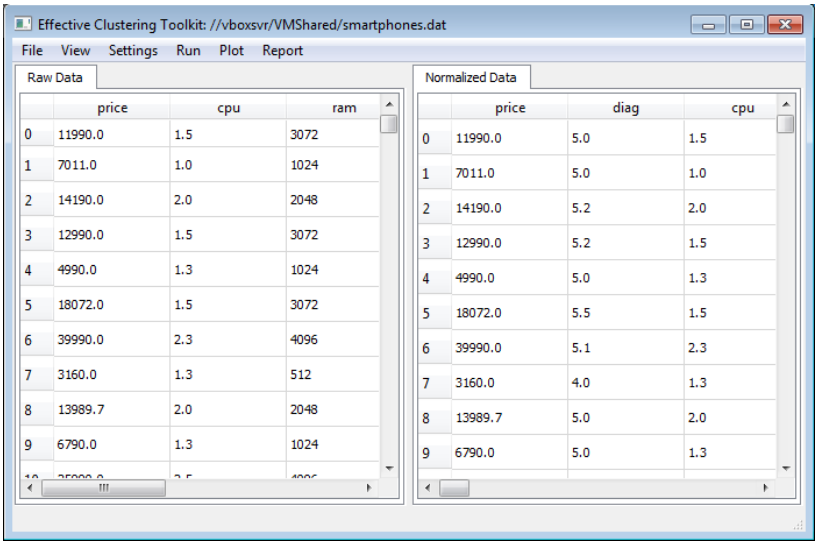
После нормализации признаков результат будет отображен во вкладке “`Normalized Data`”



6.4 Отбор признаков

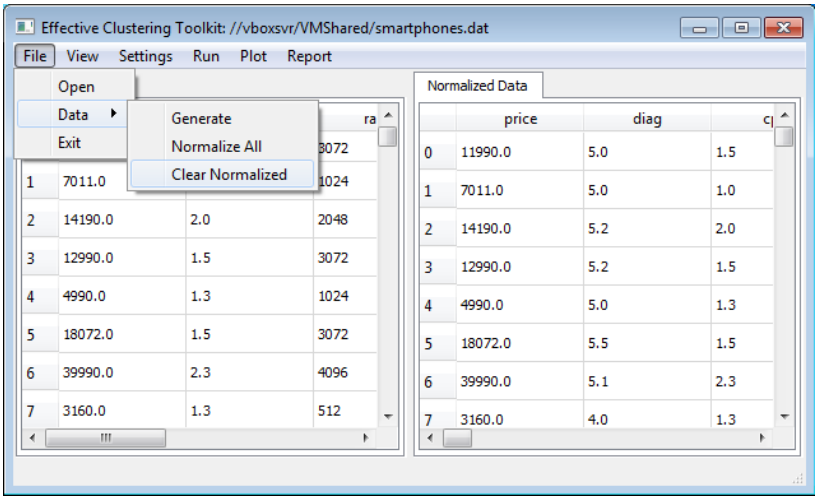
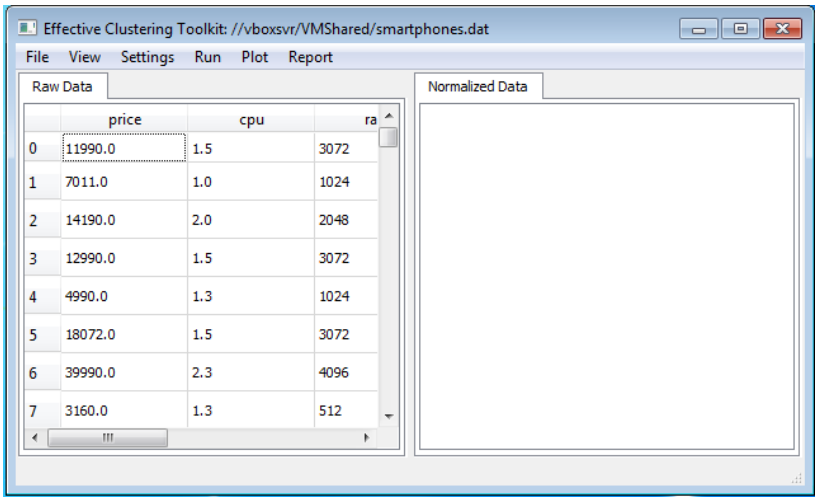
6.4.1 Удаление одного признака

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.	

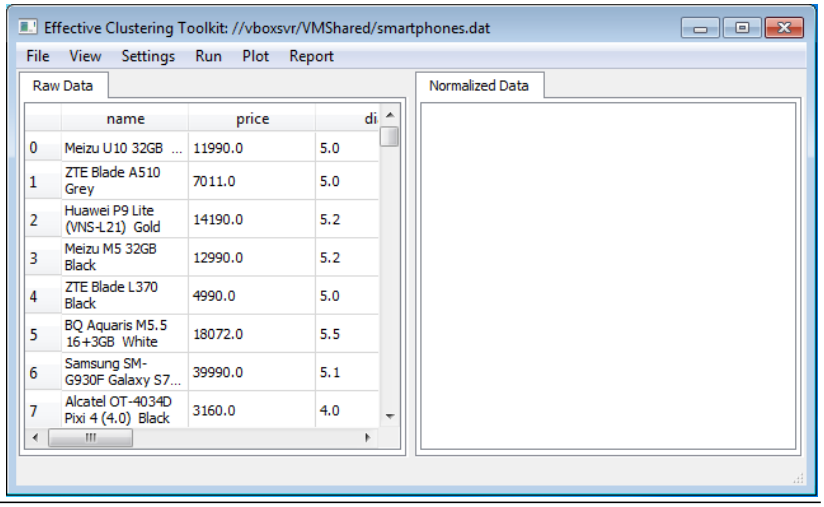
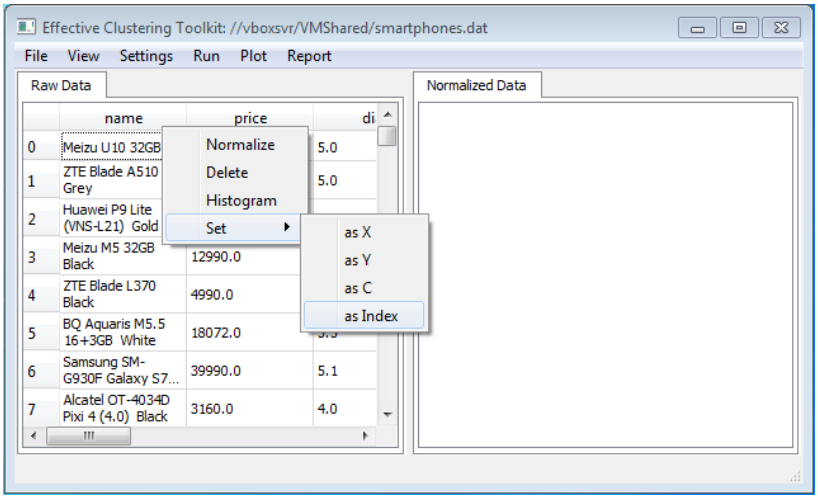
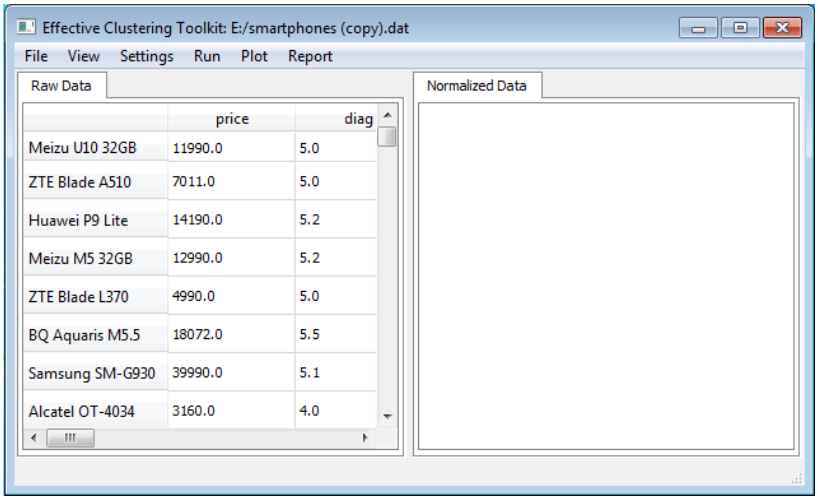
6.4.2 Удаление всех признаков сразу

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.	

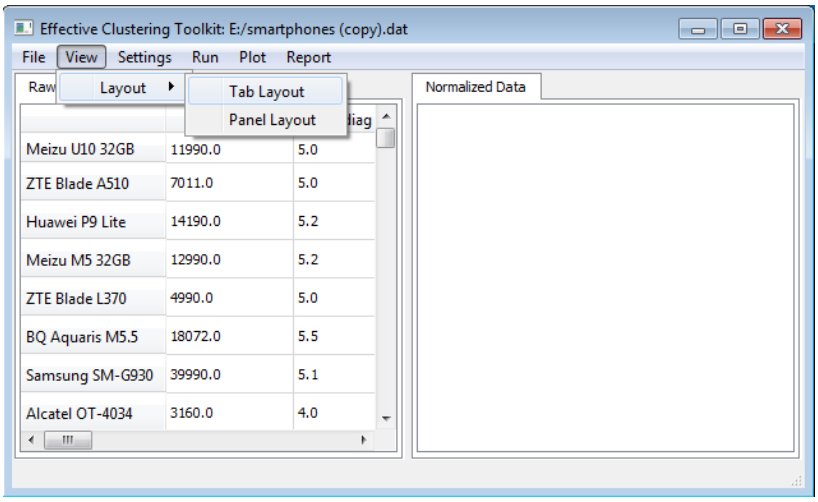
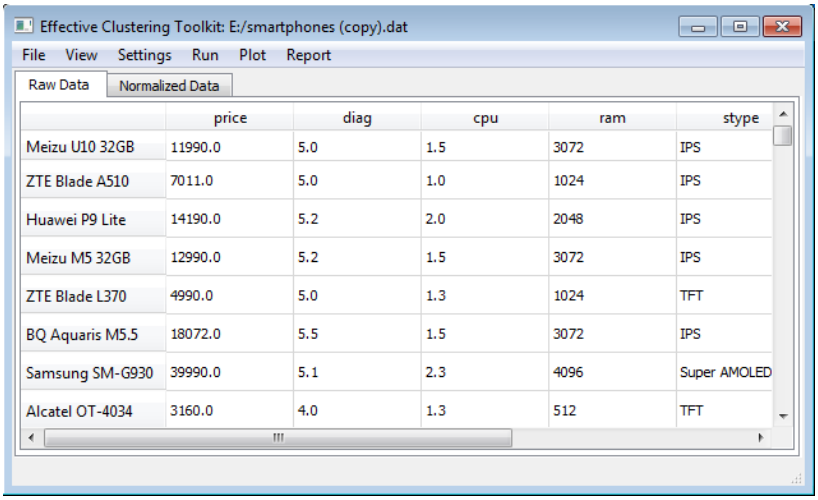
6.4.3 Установка признака как индекс

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 “6.2 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`	
3. Посмотреть результат Результат выполнения предыдущего действия состоит в отображении признака в колонке индекса как показано на рисунке справа.	

6.5 Настройка способа отображения вкладок

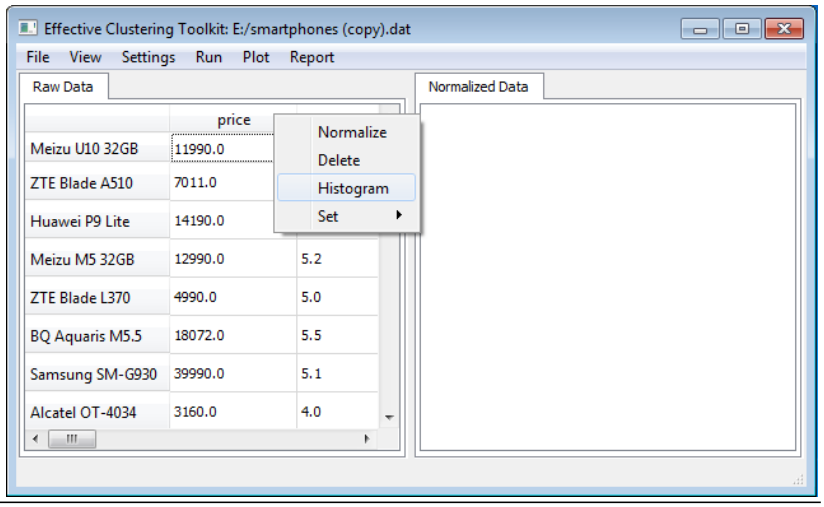
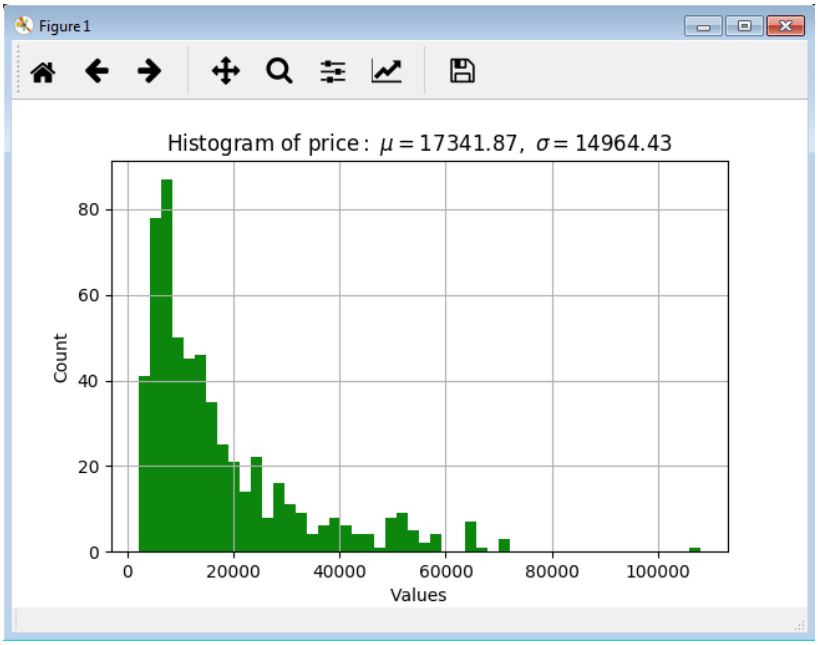
Программа имеет два способа отображения данных: с помощью вкладок и с помощью панелей (по умолчанию).Для переключения этих способов служит пункт меню `View`

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.	

6.6 Визуализация

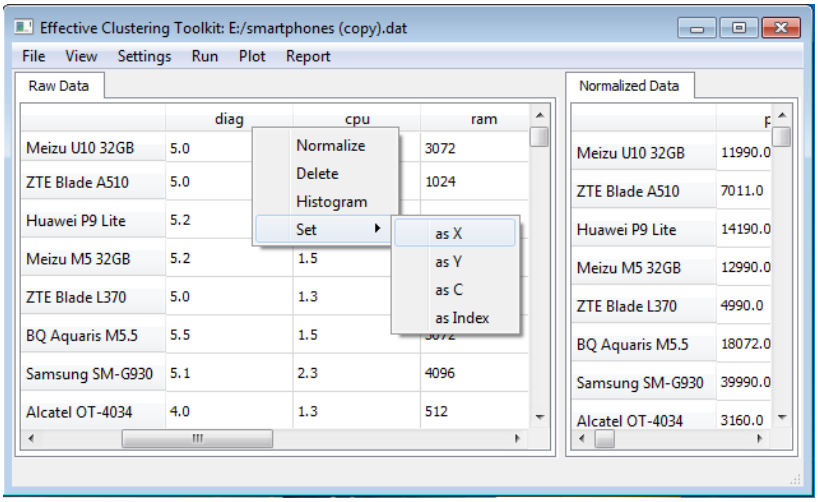
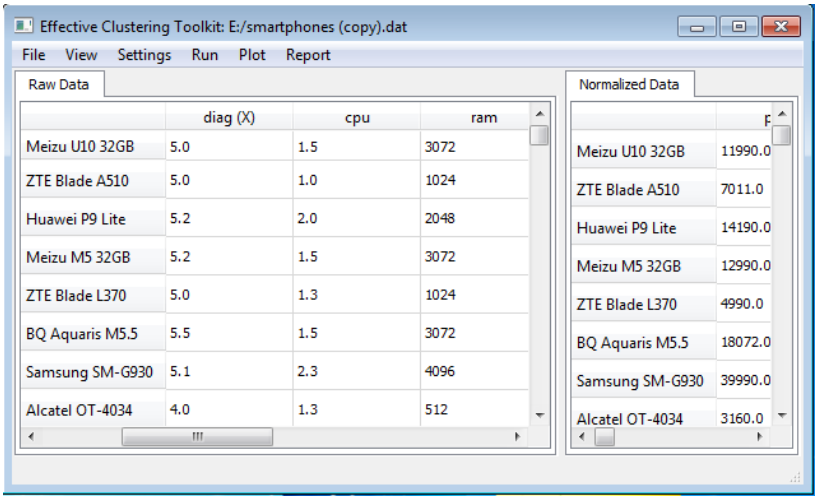
6.6.1 Построение гистограммы по признаку

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.	

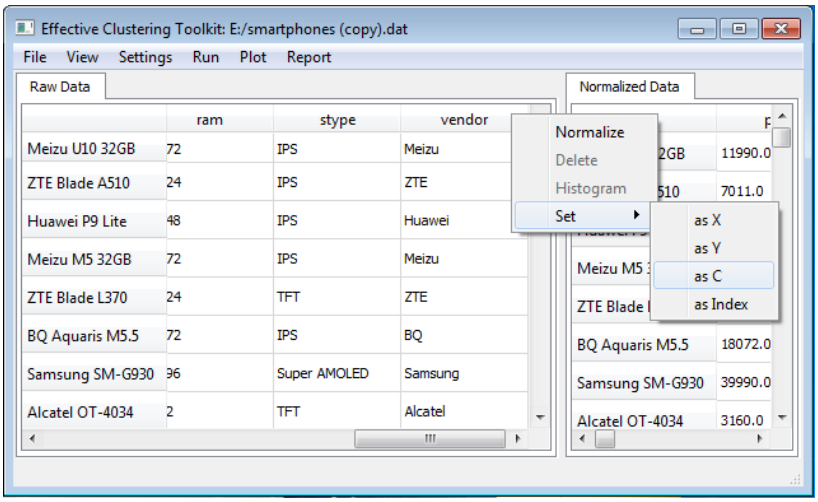
6.6.2 Построение поля рассеяния (scatter plot)

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 признак добавиться “(X)”	
3. Выбрать признак по оси Y	Аналогично пунктам 1,2.

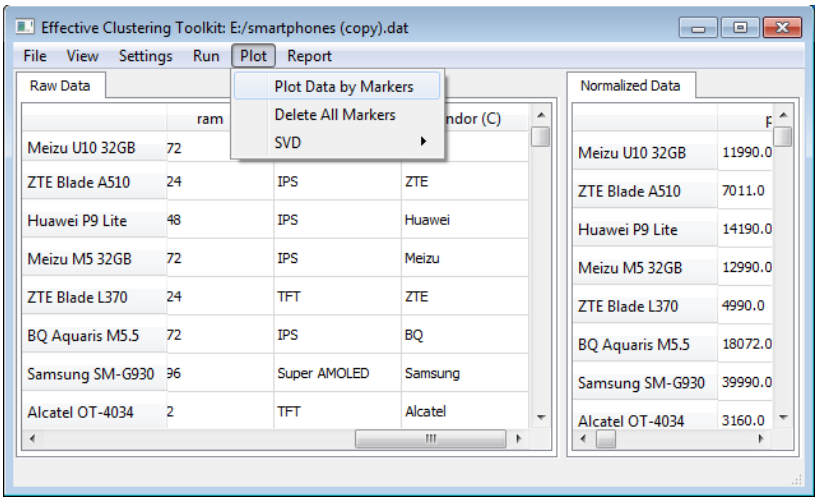
4. Выбрать признак отвечающий за цвет точек

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`



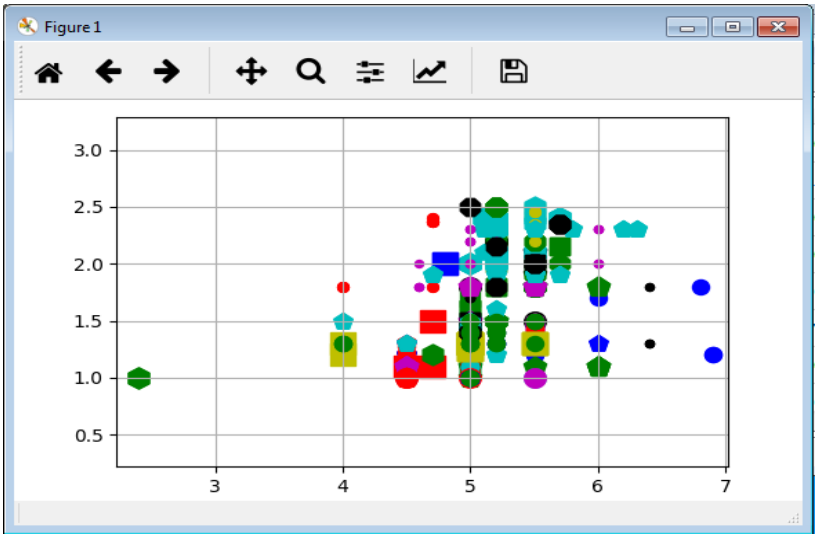
5. Построить sctter plot

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



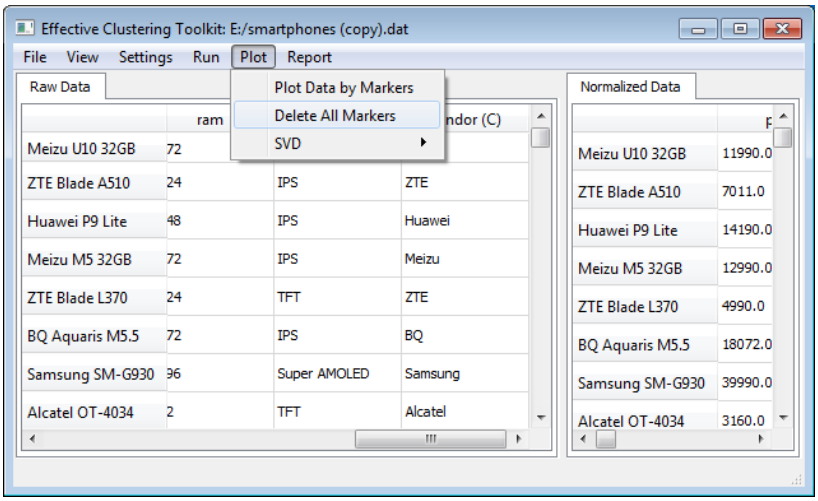
6. Посмотреть результат

В новом окне откроется построенная диаграмма



7. Удалить метки (опционально)

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



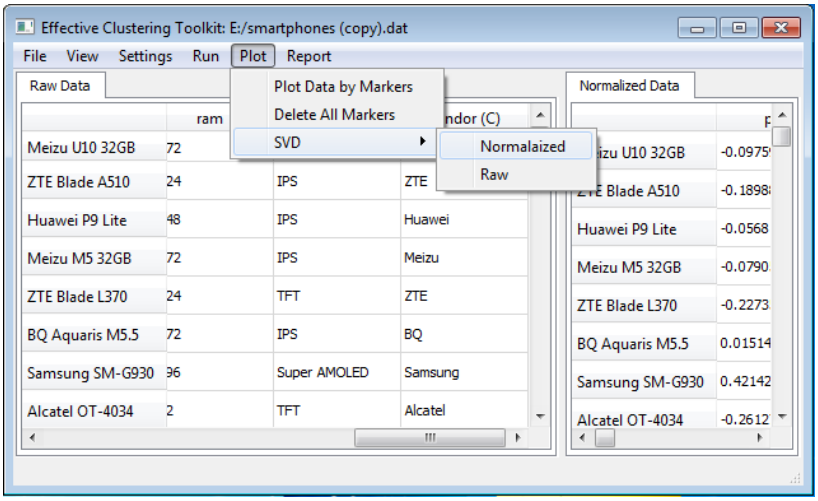
6.6.3 Построение SVD диаграммы

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

Действие/Описание	**Интерфейс**

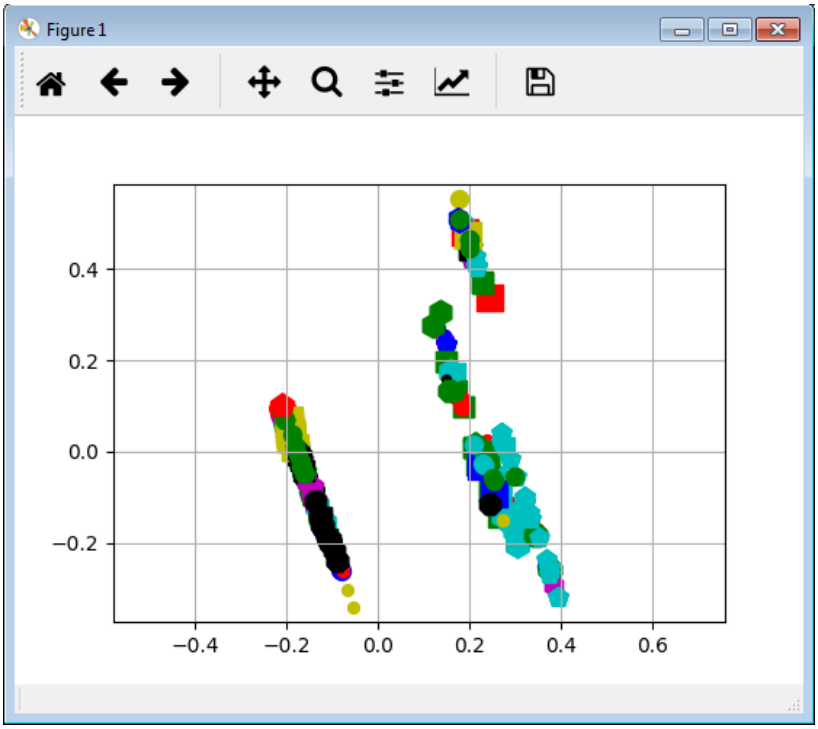
1. Построить SVD диаграмму

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



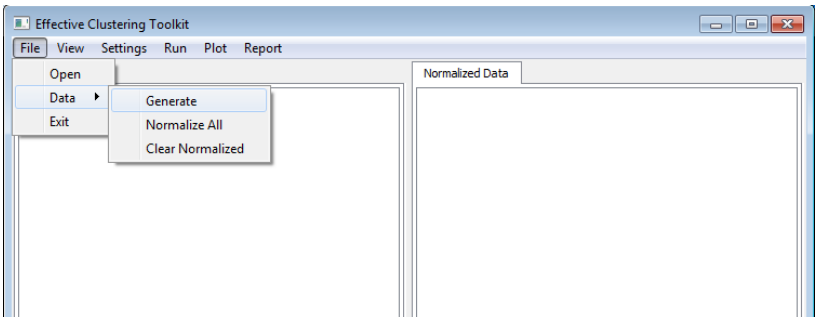
2. Посмотреть результат

Построенная диаграмма откроется в новом окне.



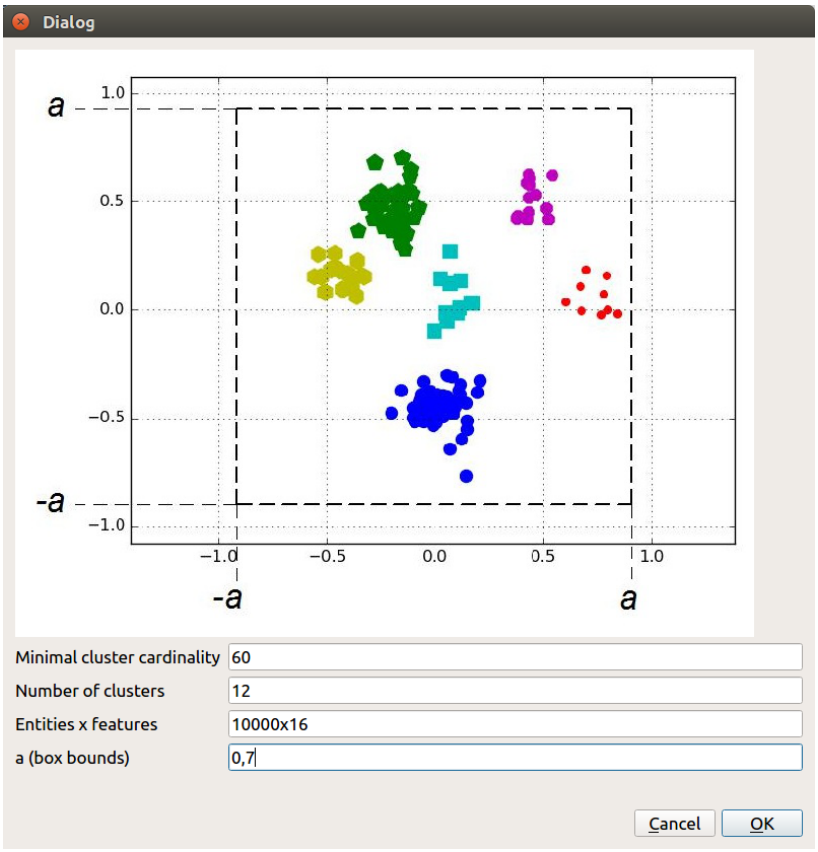
6.7 Генерация синтетических данных

Для генерирования искусственных данных необходимо вызвать диалог настройки параметры, 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.

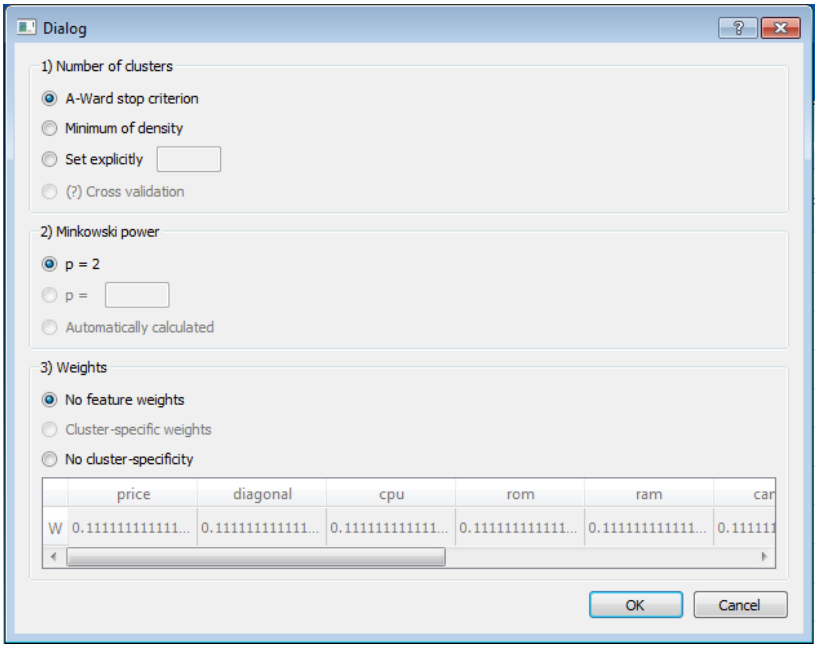
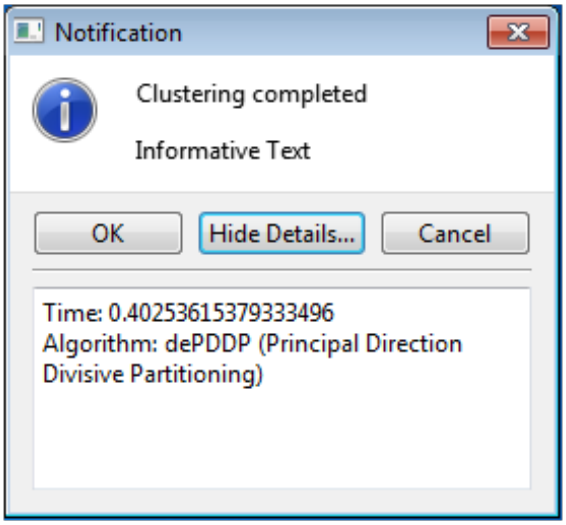


3. Сохранить данные в файл

Сохранить данные в файл в стандартном файловом диалоге.

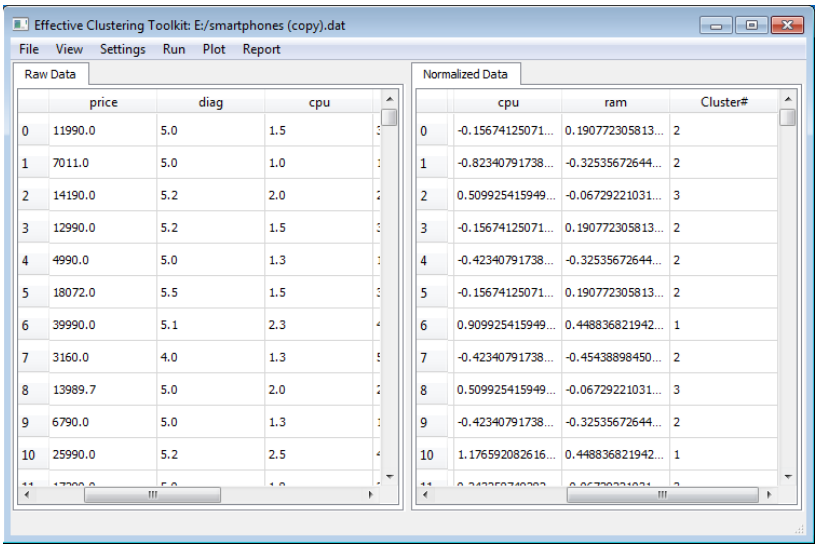
6.8 Запуск кластеризации

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 7 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 ставиться в соответствие номер кластера, которому он принадлежит. Для заданного объекта этот номер можно посмотреть в столбце “`Cluster#`”



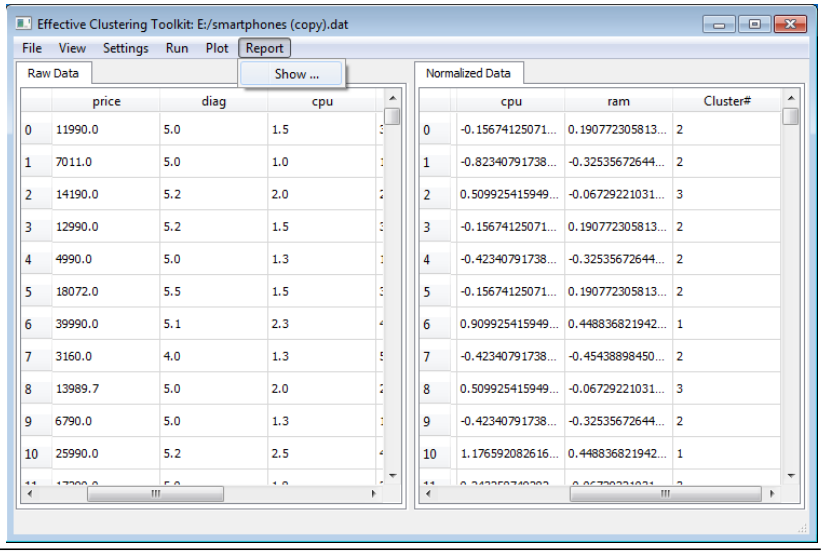
6.9 Генерация отчёта

Результаты кластеризации удобно анализировать по сгенерированному отчёту.

Действие/Описание	**Интерфейс**

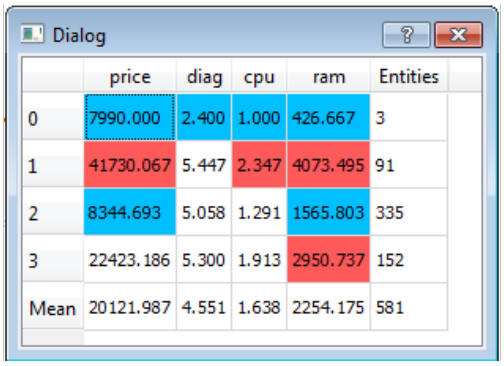
1. Сгенерировать отчёт

Для генерации отчёта в главном меню выбрать `Report` ⇒ `Show`.



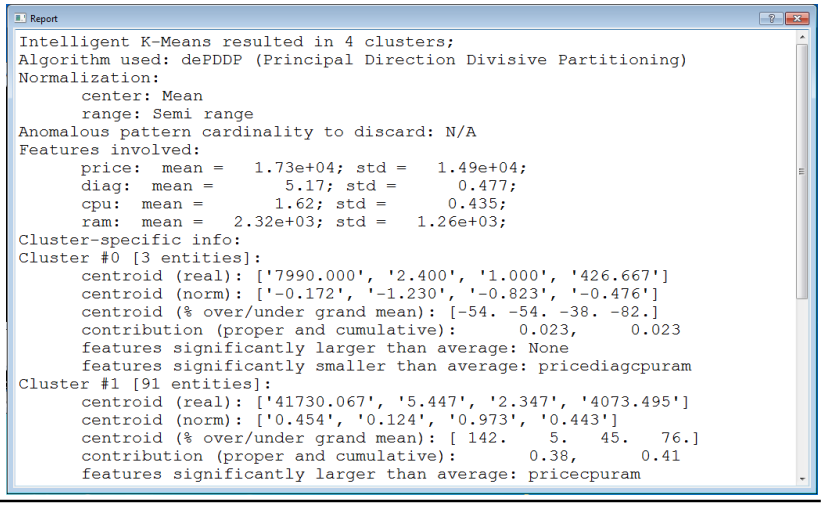
2. Посмотреть окно таблицы результатов

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.

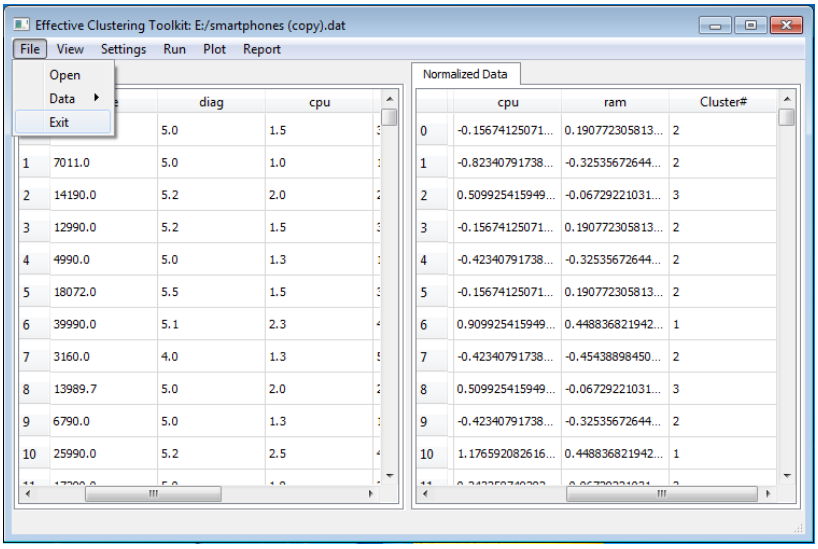


3. Посмотреть окно текстового отчёта

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



6.10 Выход из программы

Действие/Описание	Интерфейс
1. Выйти из программы Для выхода из программы в главном меню выбрать `File` ⇒ `Exit`.	

7 Алгоритмы кластеризации (краткое описание)

7.1 Алгоритм $A - Ward$

Алгоритм A-Ward является усовершенствованием широко известного алгоритма иерархической кластеризации Ward [5]. Этот алгоритм основан на пошаговом слиянии двух ближайших кластеров. На первом шаге все кластеры представлены в виде синглтонов, то есть кластеров, состоящих из единственного объекта, являющегося центроидом. Остановка алгоритма происходит при достижении числа кластеров, заданного пользователем. Для определения степени близости используется критерий, учитывающий число объектов в кластере и расстояние между центроидами.

Существенный недостаток алгоритма Ward — его время работы. Для кластеризации большого числа объектов на начальных этапах работы требуется осуществить попарный перебор большого числа синглтонов. Авторы статьи [1] предлагают устранить этот недостаток при помощи предварительного разбиения объектов. Предварительное разбиение используется как начальное состояние для работы Ward. В алгоритме Ward предлагается начинать рассмотрение с простейших кластеров, состоящих из единственного объекта, в то время как A-Ward использует кластеры, полученные при помощи метода аномальной кластеризации.

Метод аномальной кластеризации находит и удаляет аномальные кластеры по одному, начиная со всего набора данных, до тех пор пока не останется объектов для рассмотрения. В основе этого метода лежит критерий k-means [6]. Аномальным называется такой кластер, который наиболее удалён от центра данных. Центроид аномального кластера обновляется на каждом шаге, в то время как центр данных остаётся неизменным.

7.2 Алгоритм $A - Ward_{p\beta}$

В реальных приложениях требуется анализировать зашумленные данные, включающие нерелевантные признаки. В этом случае оба алгоритма Ward и A-Ward проявляют себя как непригодные. Снизить влияние нерелевантных признаков позволяет введение весовых коэффициентов. В процессе работы алгоритма для каждого признака вычисляется вес, обратно пропорциональный разбросу признака внутри кластера.

Таким образом, алгоритм A-Ward$_{p\beta}$ можно рассматривать как обобщение алгортма A-Ward. Основная идея обобщения состоит в том чтобы использовать метрику Минковского произвольной степени, а также назначить каждому признаку в пределах кластера определённый весовой коэффициент, учитывающий разброс этого признака внутри кластера. Параметры p и β являются степенями Минковского и весовых коэффициентов 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 для алгоритма A-Ward$_{p\beta}$ аномальная кластеризация обобщена с учётом новых параметров.

7.3 Алгоритм $BiKM - R$

Алгоритм bisecting k-means (раздвоение по методу k-средних) разбивает некоторый кластер S на два, при этом используя критерий минимума средних квадратов. Для инициализации алгоритма требуется указать начальные центроиды c_1 и c_2. На следующем этапе происходит альтернативная минимизация суммарной квадратичной ошибки, которая осуществляется за два шага. На первом шаге обновляется кластеры, то есть все объекты разделяются на те, которые расположены ближе к центроиду c_1 (они образуют кластер $S1$) и те, которые ближе к c_2 (кластер $S2$). На втором шаге для кластеров $S1$ и $S2$ вычисляются новые центориды c_1' и c_2'. Процесс повторяется до тех пор пока происходят изменения центроидов кластеров.

Результат алгоритма раздвоения может сильно зависеть от выбора исходных центроидов. Как правило, рекомендуется случайный выбор, но такое решение обеспечивает нестабильный результат. Как и в случае с агломеративным алгоритмами, метод аномальных кластеров может быть использован для предварительного выявления центроидов рациональным образом. В такой постановке для инициализации алгоритма раздвоения используются центроиды двух наибольших аномальных кластеров.

Для остановки алгоритма авторами статьи [4] предложен новый критерий, основанный на проецировании точек кластеров на произвольные направления. Пусть на некотором этапе работы алгоритма имеется S_k кластеров. Первым шагом генерируются s случайных векторов p_i, $i = 1, ..., s$. Для генерации используется нормальное сферическое распределение со средним в начале координат и $\sigma^2 = \frac{1}{V}$, где V – количество признаков. На втором шаге каждый элемент x каждого кластера S_k $(k = 1, ..., K)$ проецируется на направления p_i, координаты проекции определяются как скалярное произведение: $x_t =< x, p_t >$. Для каждого направления вычисляется функция плотности f_k^t по методу ядерной оценки. Если для некоторого кластера k отношение ϵ_k числа направлений, для которых функции плотности f_k^t имеют по крайне мере один минимум к общему числу направлений меньше заданного пользователем порога ϵ , то кластер k не разбивается. Для разделения выбирается в первую очередь кластер с наибольшим отношением ϵ_k/ϵ . Выбранный кластер разбивается но наиболее глубокому минимуму.

7.4 Алгоритм $dePDDP$

Алгоритм dePDDP (Principal Direction Divisive Partitioning) относится к иерархическим дивизивным и является модифицированной версией алгоритма PDDP [7] . Изначально критерий разделения кластера на две части был относительно простым: предлагалось разделить кластер по главной компоненте на положительную и отрицательную части. Впоследствии эта идея была усовершенствована при помощи правила, учитывающего распределение данных [8]. Разбиение производится по наиболее глубокому минимуму функции плотности данных, спроецированных на первую главную компоненту. Это правило примечательно тем, что может быть использовано для разрешения двух сопряженных 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,

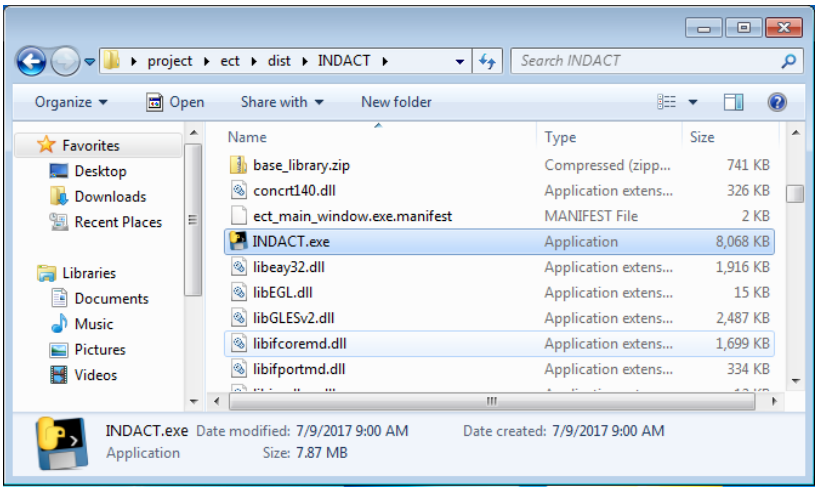
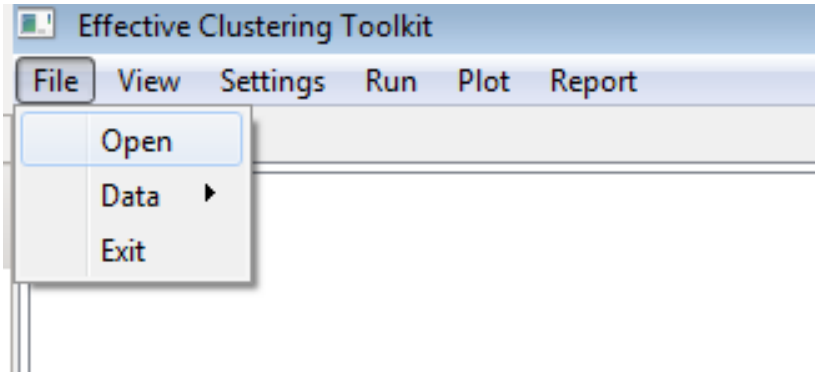
то такой кластер не может быть разделен по критерию данного алгоритма. Экспериментально было показано, что алгоритм, работающий на описанных принципах эффективно решает задачу кластеризации как на реальных данных, так и на синтетических. Оценка функции плотности осуществляется по методу ядерной оценки.

8 Примеры работы с программой

8.1 Нормализация

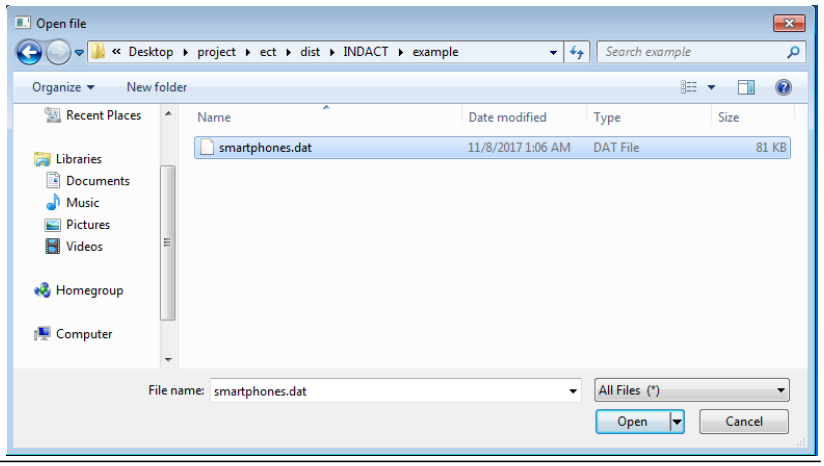
8.1.1 Нормализация с центрированием по среднему и масштабированием по полуразмаху

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`.	

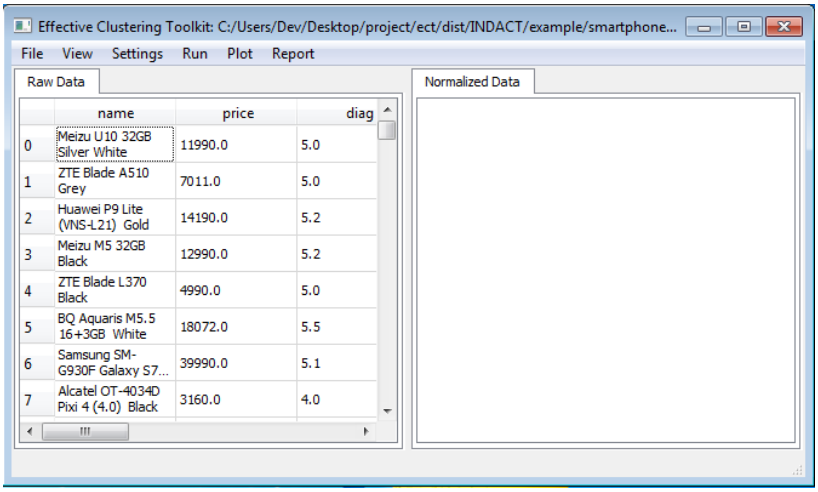
3. Выбрать текстовый файл с данными

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



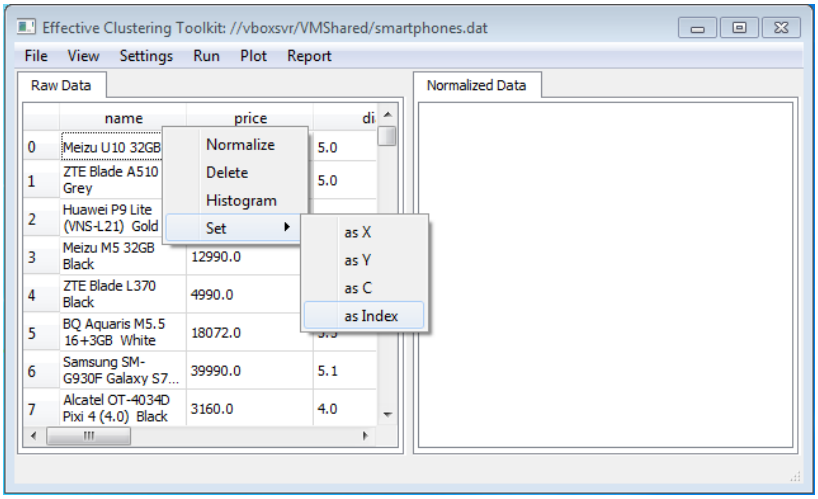
4. Убедиться что файл загружен

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



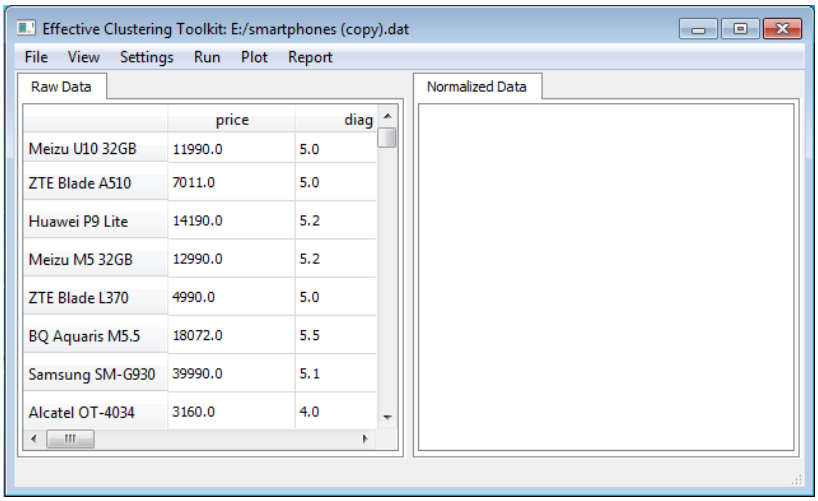
5. Установить признак `name` как индекс

Отрыть контекстное меню ПКМ и выбрать `Set` ⇒ `As Index`



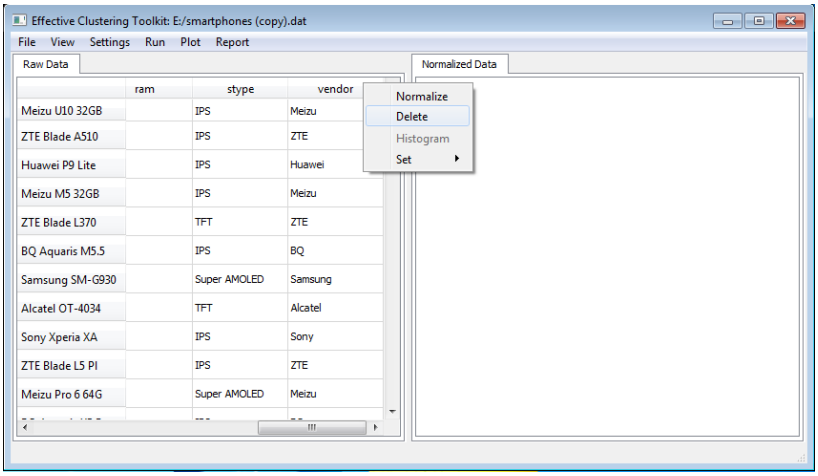
6. Убедиться что признак выставился

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



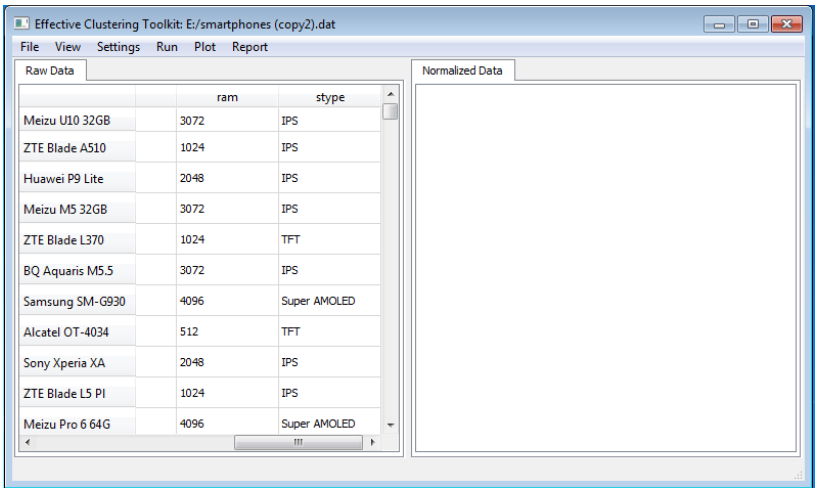
7. Удалить признак `vendor`

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



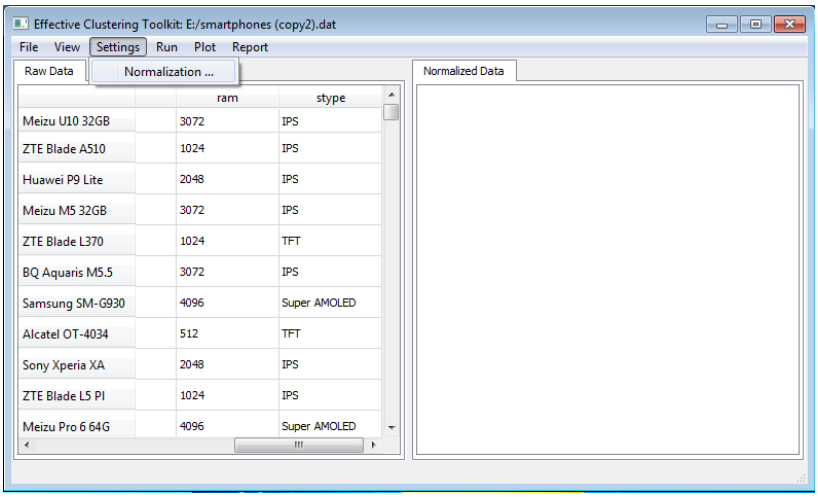
8. Убедиться что признак удалён

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



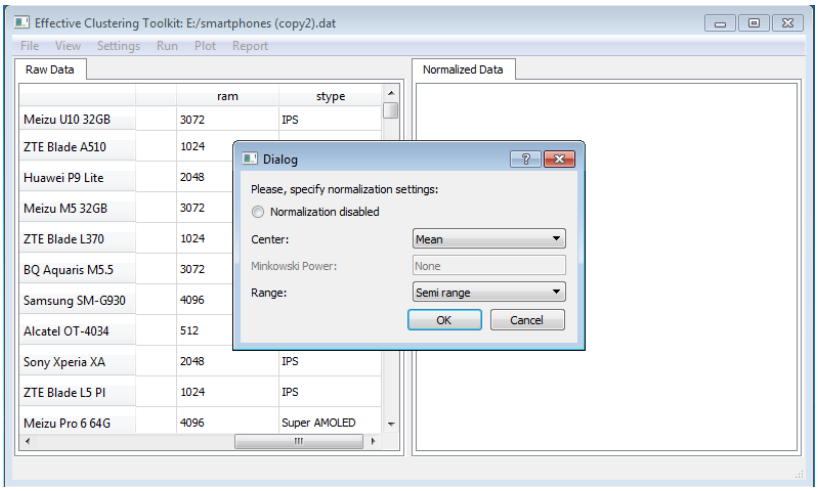
9. Открыть окно нормализации

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



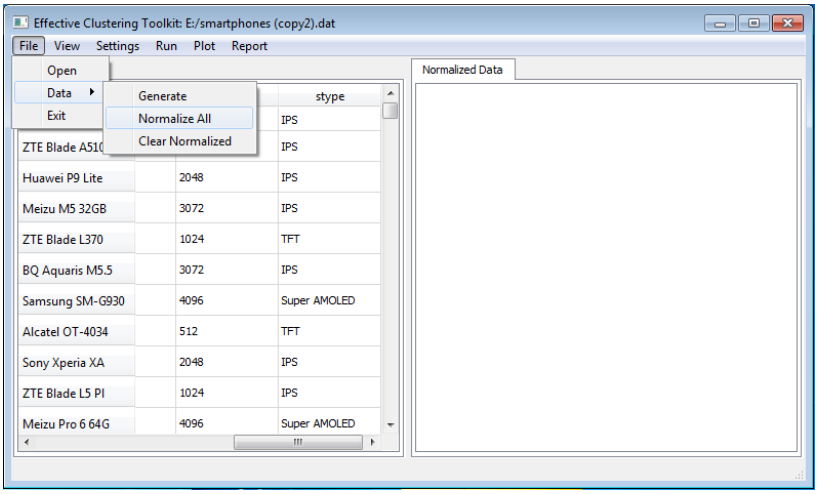
10. Выставить параметры нормализации

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`.



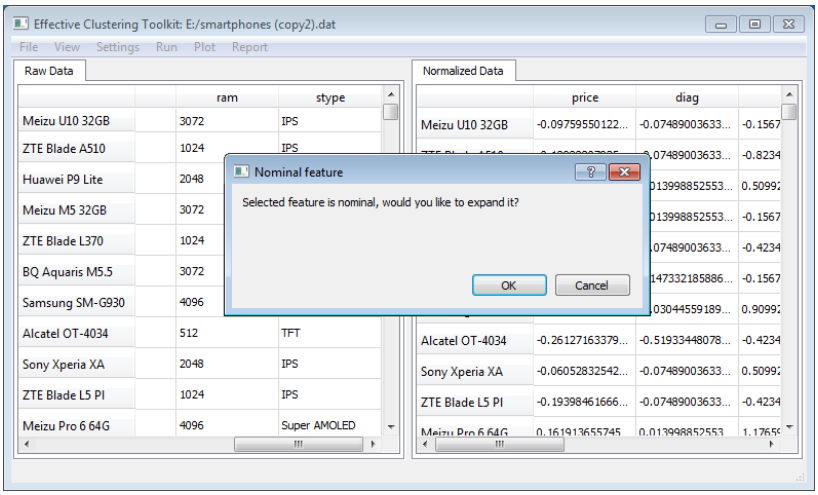
11. Запустить нормализацию всех признаков

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



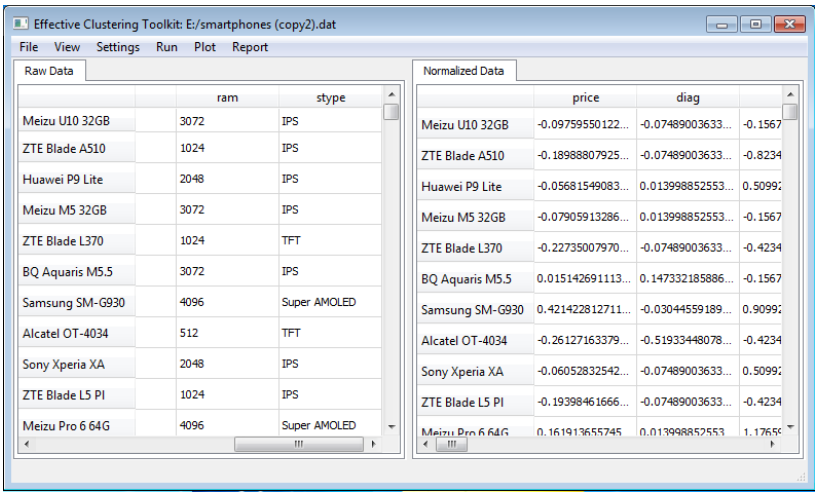
12. Подтвердить нормализацию категориального признака

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`”



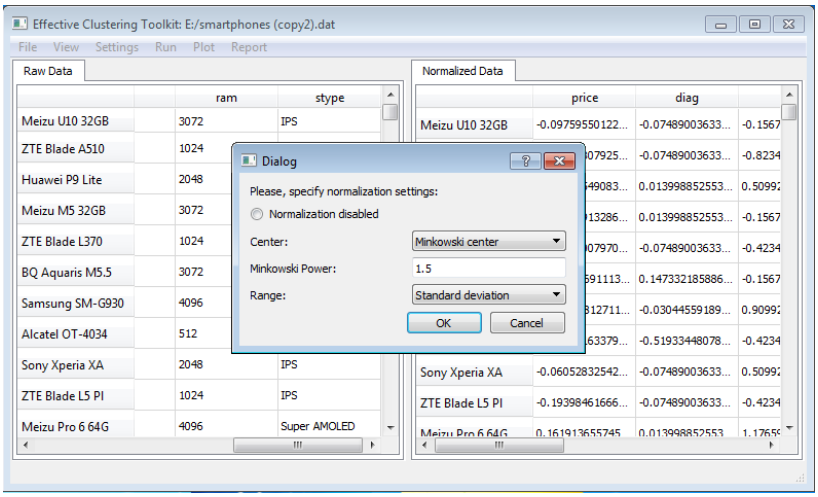
8.1.2 Нормализация с центрированием по Минковскому и масштабированием по стандартному отклонению

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 (8.1.1).

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`.



---

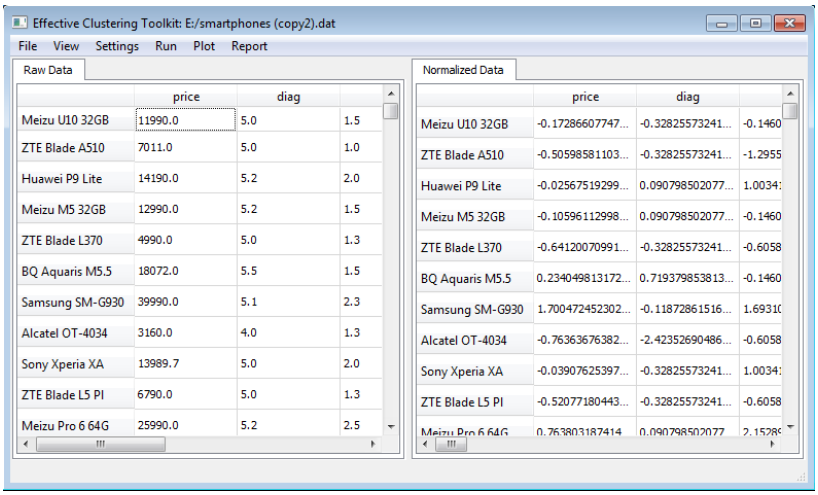
3. Нормализовать все признаки

Выполнить пункты 11–12 из предыдущего примера (8.1.1).

---

4. Посмотреть результат

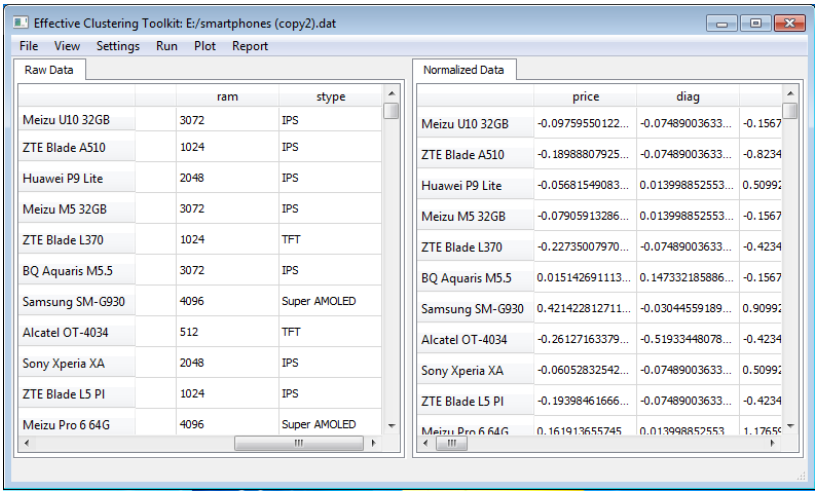
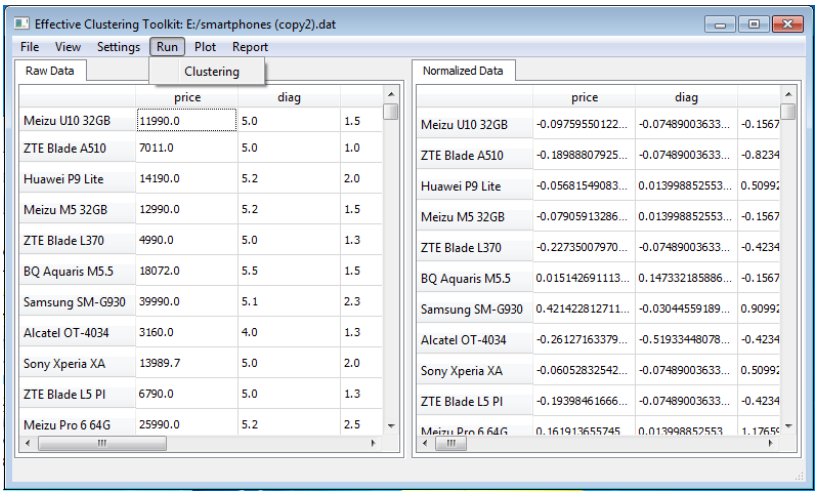
После нормализации признаков результат будет отображён во вкладке “`Normalized Data`”



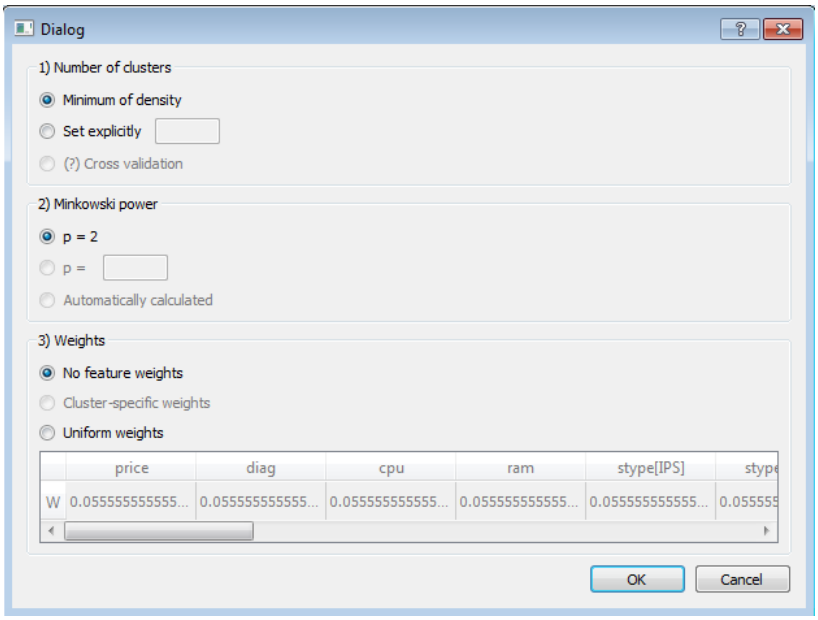
8.2 Кластеризация

8.2.1 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. 8.1.1).

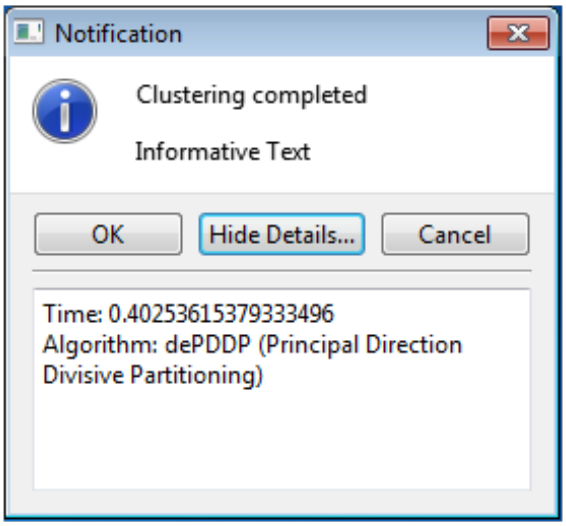
Действие/Описание	Интерфейс
1. Запустить программу, загрузить файл и нормализовать признаки Выполнить все пункты из первого примера(8.1.1). Для кластеризации требуются нормализованные признаки.	
2. Открыть окно кластеризации Пред запуском кластеризации потребуется задать общие параметры процедуры, по которым программа выберет конкретный алгоритм. Для этого необходимо из главного меню выбрать: `Run` ⇒ `Clustering`.	

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`.

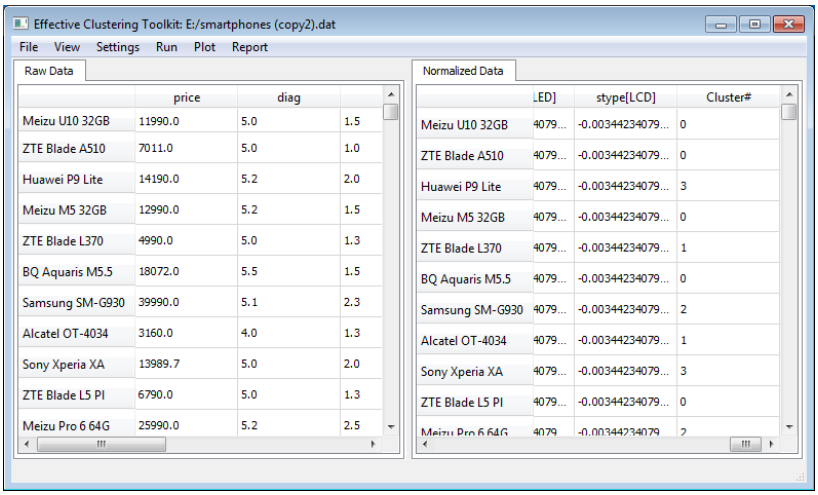
4. Дождаться завершения кластеризации



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`.

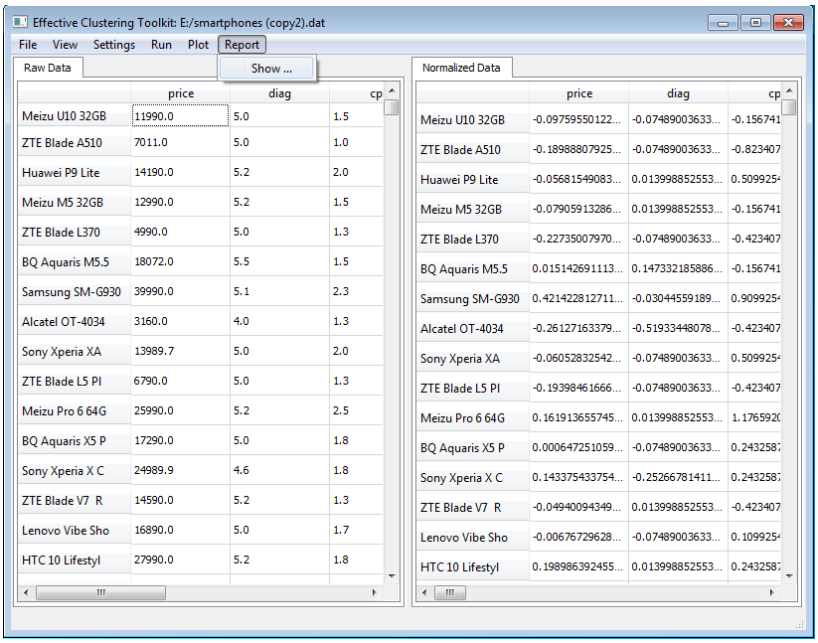
5. Посмотреть кластерную принадлежность

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



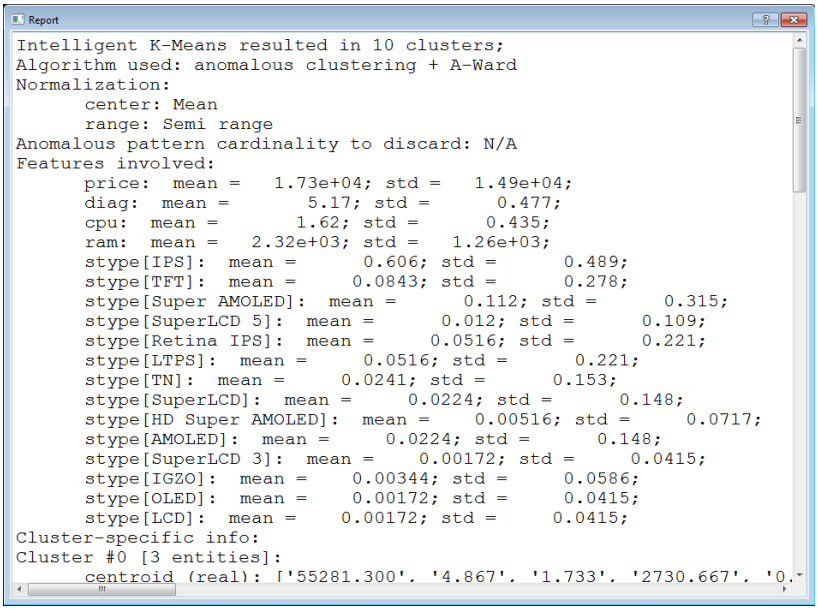
6. Сгенерировать отчёт

Для генерации отчёта в главном меню выбрать `Report` ⇒ `Show` (см. 6.9).



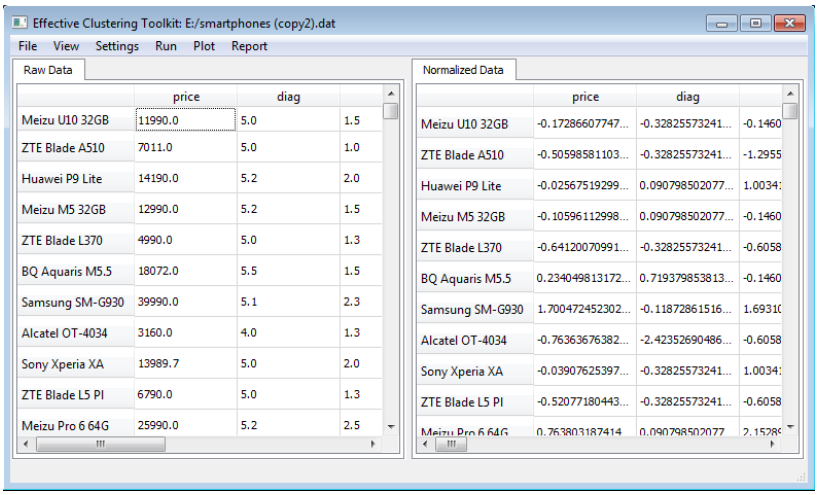
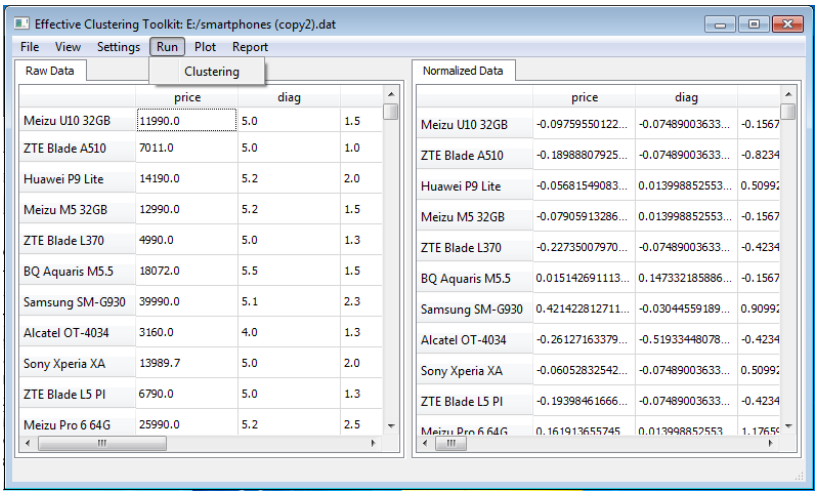
7. Посмотреть отчёт

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

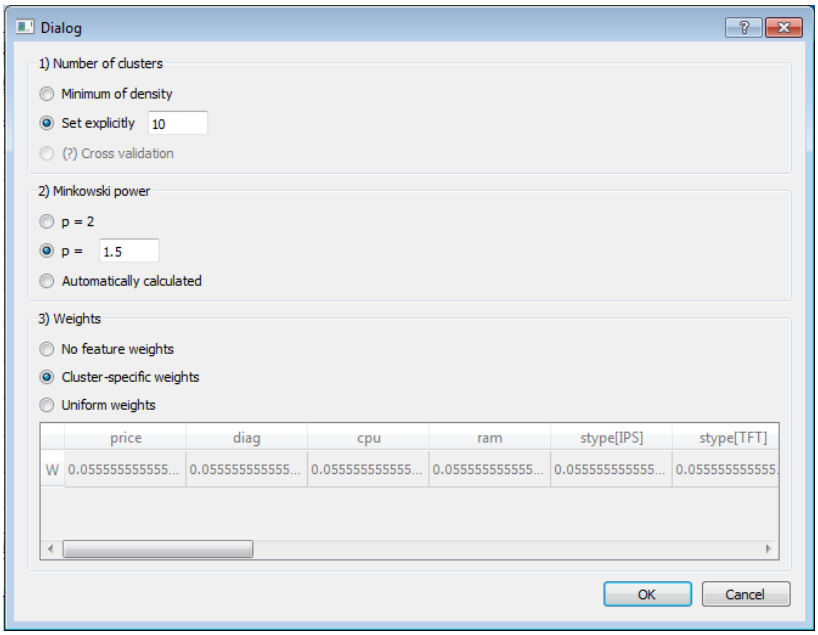


8.2.2 Кластеризация с явно заданным числом кластеров

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 7.1).

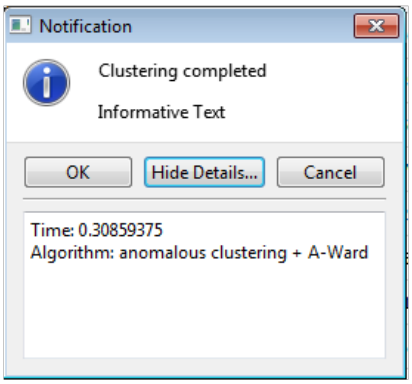
Действие/Описание	Интерфейс
1. Запустить программу, загрузить файл и нормализовать признаки Выполнить все пункты из второго примера (8.1.2). Для кластеризации требуются нормализованные признаки. Для данного примера используется нормализация с центрированием Минковского.	
2. Открыть окно кластеризации Пред запуском кластеризации потребуется задать общие параметры процедуры, по которым программа выберет конкретный алгоритм. Для этого необходимо из главного меню выбрать: `Run` ⇒ `Clustering`.	

3. Установить параметры алгоритма кластеризации



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`.

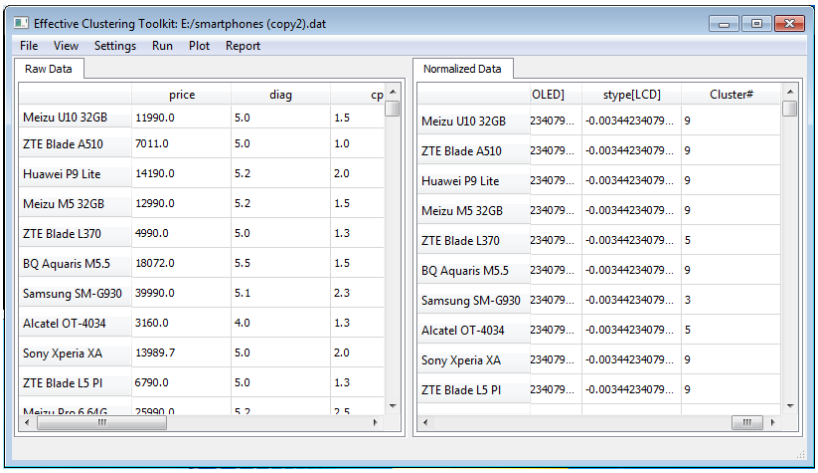
4. Дождаться завершения кластеризации



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`.

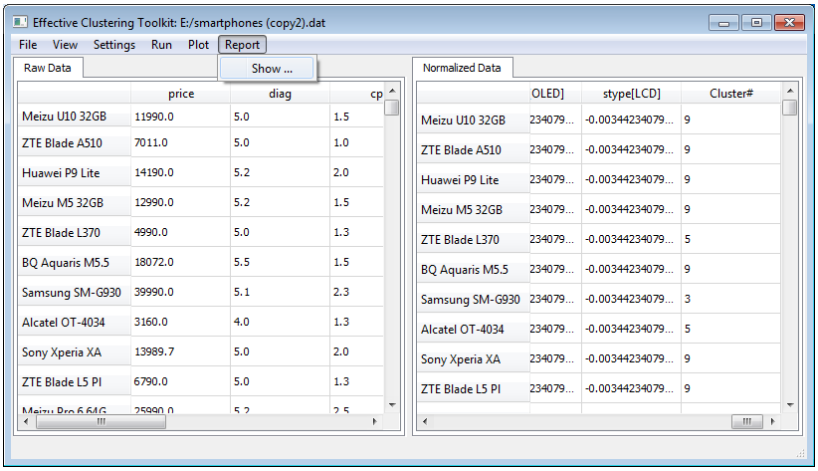
5. Посмотреть кластерную принадлежность

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



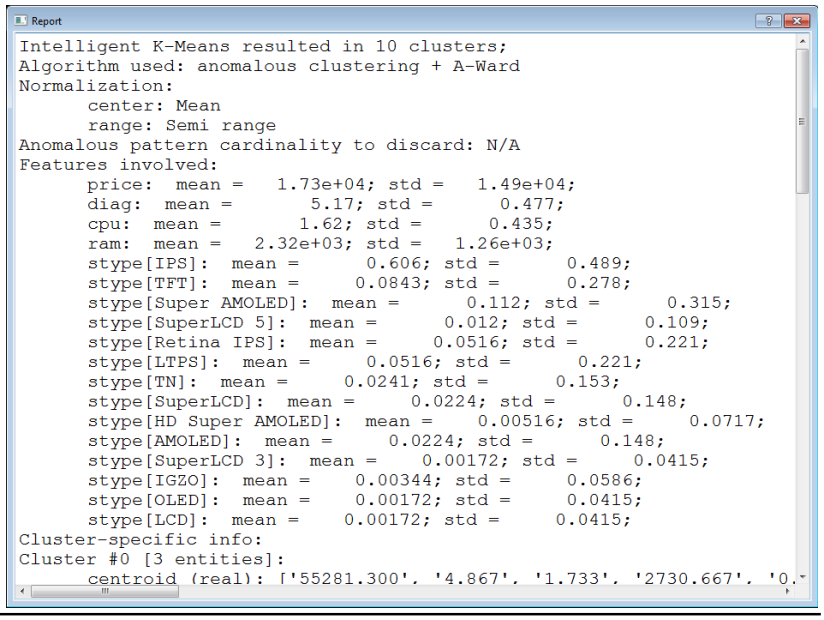
6. Сгенерировать отчёт

Для генерации отчёта в главном меню выбрать `Report` ⇒ `Show` (см. 6.9).



7. Посмотреть отчёт

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.



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 INDUCT.

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. 1.

диапазон нормирования : величина в знаменателе формулы 1.

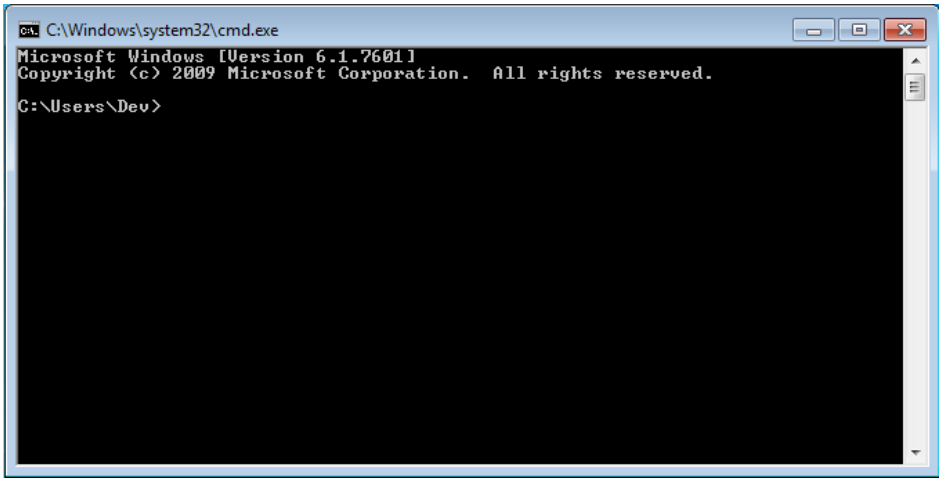
исходные данные : набор данных, представленных в виде плоской таблице и сохранённой в текстовом файле. Строка таблицы соответствует одному объекту, а столбец — признаку. Пример таблицы исходных данных показан в разделе 5.3.

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. 1.

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 // Jornal of American Statical 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.